METHODS IN CREATING ALTERNATE ASSESSMENTS: CALIBRATING A MATHEMATICS ALTERNATE ASSESSMENT DESIGNED FOR STUDENTS WITH DISABILITIES USING GENERAL EDUCATION STUDENT DATA

by

EUNJU JUNG

A DISSERTATION

Presented to the Department of Special Education
and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2008

University of Oregon Graduate School

Confirmation of Approval and Acceptance of Dissertation prepared by:

EunJu Jung

Title:

"Methods in Creating Alternate Assessments:Calibrating a Mathematics Alternate Assessment Designed for Students with Disabilities Using General Education Student Data"

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Elizabeth Harn, Co-Chairperson, Special Education and Clinical Sciences
Paul Yovanoff, Co-Chairperson, Educational Leadership
Gerald Tindal, Member, Educational Leadership
Robert O Brien, Outside Member, Sociology

and Richard Linton, Vice President for Research and Graduate Studies/Dean of the Graduate School for the University of Oregon.

December 13, 2008

Original approval signatures are on file with the Graduate School and the University of Oregon Libraries.

An Abstract of the Dissertation of

Eunju Jung for the degree of Doctor of Philosophy

in the Department of Special Education and Clinical Sciences

to be taken December 2008

Title: METHODS IN CREATING ALTERNATE ASSESSMENTS: CALIBRATING A MATHEMATICS ALTERNATE ASSESSMENT DESIGNED FOR STUDENTS WITH DISABILITIES USING GENERAL EDUCATION STUDENT DATA

Approved: ______________________________
Dr. Beth Harn, Co-Chair

Approved: ______________________________
Dr. Paul Yovanoff, Co-Chair

The NCLB (2001) requires states to test all students from grades 3 through 8, including students with disabilities, whom the tests were not designed to assess. This study focused on students with disabilities referred to as '2%' students. Two percent students are students with disabilities for whom the regular state assessment is considered too difficult, yet the alternate academic achievement standards are too easy.

A significant challenge in developing alternate assessments is obtaining suitable sample sizes. This study investigated whether psychometric characteristics of mathematic alternate assessment items created for 2% students in grade 8 can be meaningfully estimated with data obtained from general education students in lower grades. Participants included 23 2% students in grade 8 and 235 general education students in grades 6-8. Twenty three 2% students were identified through the Student Performance Test (10 standard items and 10 2% items) and the Teacher Perception Survey. Performance on 10 2% items by the 2% students and the general education students were analyzed to address the questions: (a) are there grade levels at which the item parameters estimated from general education students in grade 6-8 are not different from those obtained using the 2% student sample?; and (b) are there grade levels at which the estimated ability of general education students in grades 6-8 are not different the 2% student sample in grade 8?

Results indicated that the item response patterns of 2% students in grade 8 were comparable to those of general education students in grades 6 and 7. Additionally, 2% students in grade 8 showed comparable mathematics performance on 2% items when compared to general education students in grades 6 and 7. Considering the content exposure of students in lower grades, this study concluded that data from general education students in grade 7 would be more appropriate to be used in designing alternate assessment for 2% students in grade 8 than data from students in grade 6. The general conclusion is that using data obtained from general education students in lower grade levels may be an appropriate and efficient method of designing alternate assessment items.

CURRICULUM VITAE

NAME OF AUTHOR: Eunju Jung

PLACE OF BIRTH: Pusan, Korea

DATE OF BIRTH: October 13, 1974

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon
Seoul National University, Seoul, Korea
Pusan National University, Pusan, Korea

DEGREES AWARDED:

Doctor of Philosophy in the Department of Special Education and Clinical Sciences, 2008, University of Oregon
Master of Arts in Special Education, 2004, Seoul National University
Bachelor of Arts in Special Education (Summa Cum Laude), 2001, Pusan National University
Bachelor of Science in Mathematics, 1997, Pusan National University

AREAS OF SPECIAL INTEREST:

Mathematics assessment and instruction for students with disabilities
Large-scale assessment development and validation
Methodology including statistical and quantitative methods

PROFESSIONAL EXPERIENCE:

Teaching Assistant in SPED 660: Design of Instruction, Special Education, University of Oregon, Eugene, Oregon, 2008

Research Assistant, Behavioral Research and Teaching, University of Oregon, Eugene, Oregon, 2006-2008

Teaching Assistant in EDLD 610: Foundations of Educational Research II, Education Leadership, University of Oregon, Eugene, Oregon, 2007

Research Assistant, Center on Teaching and Learning, University of Oregon, Eugene, Oregon, 2004-2006

Research Assistant, Seoul National University, Seoul, Korea, 2003-2004

Special Educator, Seoul Jeongjin School, Seoul, Korea, 2002-2004

GRANTS, AWARDS AND HONORS:

Graduate Teaching Fellowship, University of Oregon, 2004-2008

Korean Government Overseas Scholarship, Korea, 2004-2006

Scholarship for Academic Achievement from PNU, Pusan, Korea, 2000

Governmental Scholarship for Academic Achievement from Ministry of Education, Korea, 1999

Scholarship for Academic Achievement from PNU, Pusan, Korea, 1996

Scholarship for Academic Achievement from PNU, Pusan, Korea, 1995

PUBLICATIONS:

Ketterlin-Geller, L. R., Jung, E., Geller, J., & Yovanoff, P. (in press). *Project DIVIDE instrument development*. Eugene, OR: University of Oregon, College of Education, Behavioral Research and Teaching.

Jung, E. (2008). Considerations on alternate assessments: NCLB regulations for alternate and modified academic achievement standards. *Korean Journal of Special Education, 43*(2), 123-136.

Jung, E., Liu, K., Ketterlin-Geller, L. R., & Tindal, G. (2008). *Instrument development procedures for GOM mathematics measures*. (Tech. Rep. No. 08-02). Eugene, OR: University of Oregon, College of Education, Behavioral Research and Teaching.

Kim, D., Yeo, S., Kim, K., & Jung, E. (2005). *Inclusion of special education system for the unified Korea.* Seoul, Korea: Seoul National University, College of Education, SNU Education Research.

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to the dissertation committee members: Dr. Beth Harn for her vision, guidance, and support; Dr. Paul Yovanoff for his encouragement, detailed feedback, and comforting smiles; Dr. Gerald Tindal for his clarity, thoughtful input, and continued support; and Dr. Robert O'Brien for his insights and commitment. I would also like to thank Dr. David Chard for his patience, understanding, and guidance. He inspired me to balance my doctoral study and family.

This study could not have been completed without the cooperation of students and teachers who participated. I want to thank them for their valuable contribution. Special thanks to Dr. Julie Alonzo for her significant role in data collection. This investigation was supported in part by a grant from the Institute of Education Sciences, grant number R324A07188, to Dr. Gerald Tindal at the University of Oregon.

My deepest appreciation goes out to my parents, Jongmoon Jung and Jeongsim Moon, to my sister, Juyean Jung, and to my brother, Choongyoung Jung for their endless love, trust, and encouragement.

Finally, I thank God, who has guided and blessed me.

To my husband Guisung, and my daughter Haeun
who have been with me at all phases of this journey

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Assessment is an essential procedure for determining the current academic status of students, identifying their specific needs, and designing and providing appropriate instruction. By collecting and evaluating data, we can make better decisions for students and provide them with more effective instruction (Salvia & Ysseldyke, 2004). Even though the field of education recognizes the importance of valid assessment practices to better meet the needs of students, we have not extended this same careful practice to a specific student population – students with identified disabilities. Test developers have purposely excluded students with a range of disabilities (i.e., mental retardation, autism, vision or hearing impairments, etc.) because of the challenges and complexities in test development, validation, and administration (Thurlow, Elliot, & Ysseldyke, 2003). As Almond et al. (2002) argue, no one should be excluded in educational accountability systems, and all students should benefit from their educational experiences.

Most students with disabilities may be able to take the regular assessment with or without testing accommodations. However, some students with disabilities may need alternate methods to participate in assessments. Alternate assessments benefit students with disabilities who cannot be accurately assessed by regular assessments even with testing accommodations. Attention to alternate assessment for students with disabilities has been dramatically increasing since the No Child Left Behind (NCLB) legislation was enacted in 2001. Students with disabilities may be assessed within a range of possibilities based on the nature of their unique disabilities and their relevant learning features: 1) regular state assessment, 2) regular state assessment with appropriate accommodation, 3) *alternate assessment* based on *grade-level* academic achievement standards, 4) *alternate assessment* based on *modified* academic achievement standards, and 5) *alternate assessment* based on *alternate* academic achievement standards (U. S. Department of Education, 2007). Alternate assessments enable states to provide more accurate assessments for students who cannot be accurately measured by regular state assessments with or without testing accommodations. Recently the U.S Department of Education announced regulations for assessing a small number of special education students using alternate assessments based on *modified* academic achievement standards (U.S. Department of Education, 2007). These 2007 regulations offer states flexibility in measuring the academic achievement of students with disabilities who are not covered by existing 2003 regulations. The 2003 regulations articulated *alternate* academic achievement standards for students with the most significant cognitive disabilities.

This section discusses why assessment for students with disabilities is important and why these students should be included in accountability systems. In addition, issues in developing alternate assessments for students with disabilities are presented, highlighting the need for accurate alternate assessments and the challenges in research for technical adequacy of alternate assessments.

Assessment for Students with Disabilities and Accountability Systems

The No Child Left Behind (NCLB) Act of 2001 (PL 107-110) mandated that "all" children in grades 3 through 8, including students with disabilities, must be assessed and monitored in their reading, mathematics, and science progress. In the past, students with disabilities were expected to "be" with other students in the classroom even though they received separate instruction. Now they are also supposed to show their progress on state academic content standards similar to their peers without disability.

It is reasonably argued that if students with disabilities were not included in the development of the test used in the accountability systems, the results gathered from such instruments may not provide accurate and valid results thereby skewing school interpretation and inaccurate educational decision making for the student (Thurlow, Elliotte, & Ysseldyke, 2003). In addition to legal requirements and equity concerns about educational benefit, Thurlow, Elliott, and Ysseldyke (2003) address several reasons why students with disabilities should be included in accountability systems. First, it is unlikely to get an accurate picture of education without incorporation of students with disabilities, who comprise almost 10% of all students. Students with disabilities should be included in accountability systems to make accurate comparisons among states or districts within a state. Furthermore, the authors demonstrate that the exclusion of these students from accountability systems would result in two very powerful unintended consequences: a) increased student retention rates, and b) increased referral to special education. Finally, participation of students with disabilities in assessments and inclusion in accountability systems is an essential step for "access to the general curriculum" (IDEA, 1997).

Issues in Developing Alternate Assessments for Students with Disabilities

Systematic, psychometrically sound and valid test development procedures are critical to making well informed, valid school and student-level educational decisions. Alternate assessments for students with disabilities must be mindfully developed and evaluated by generally accepted psychometric standards (Yovanoff & Tindal, 2007). However, there are challenges in the development and validation of alternate assessments.

Accurate Alternate Assessments for Valid Educational Decision

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed in the uses of test," and addresses validity as the most important issue in developing and evaluating tests. Contrary to regular state assessments designed for most public school students, alternate assessments for students with disabilities should be designed to provide more valid score interpretations for a small and unique subpopulation of students. The alternate assessments for students with disabilities should take into account the characteristics of the participating students, the knowledge and skills tested, and the design of the assessment. In addition, the purposes and uses of the alternate assessments should be articulated. In other words, alternate assessments should be evaluated in the context of the purposes of the assessment and how the results are used (e.g., instructional change) (Marion, 2007).

Tindal et al. (2003) insist that a high-quality assessment should reflect what the test intended to measure. When developing measures, test developers sometimes make two mistakes: (a) construct-irrelevant variance and (b) construct underrepresentation (Messick, 1994). Tindal et al. (2003) explain that construct-irrelevant variance refers to a factor which is not related to construct intended to measure but affects test scores unfairly. Measures with construct underrepresentation lack the depth and breadth of knowledge and skills, in which case the test does not reflect the full range of intended content. A middle school math test demonstrates these two points. Following is a mathematics item that an 8th grade student may have to solve.

> Tim was given $100 for his twelfth birthday. He's curious to see how much it will grow if he earns interest on it. His mother tells him that she has about $3000 in the same kind of account and she earned $90 last year. About how much interest could Tim expect to earn in a year?

This item is designed to assess students' skills in algebra. However, to solve this problem a student must also have strong reading, vocabulary and comprehension skills. The issue of a student's reading ability becomes construct-irrelevant variance, especially for students who have difficulty in reading. Construct-irrelevant variance also includes inappropriate test administration and not considering the special needs of students with disabilities (Haladyna, 2002).

If testing is intended to measure mathematics achievement for students in 8th grade, the measure should cover contents of algebra, data analysis and algebra, and geometry and measurement (Department of Oregon Education, 2007). Only focusing on basic knowledge and skills in algebra can create a risk of underrepresentation of the construct. Developing alternate assessments for students with disabilities may have this unintended negative result due to the lower expectations of these students (Marion, 2007). To avoid this mistake, the domain of mathematical problems should be carefully arranged in taxonomy which is also cautiously linked to required knowledge and skills (Haladyna, 2002).

All assessments, including alternate assessments, should be developed with these two potential flaws in mind, (a) construct-irrelevant variance and (b) construct underrepresentation. For students with disabilities, these measurement flaws are further complicated by performance related disabilities. Assessments developed without such considerations will result in a negative bias using a student's disability against them (Tindal et al., 2003).

Challenges in Research for Technical Adequacy of Alternate Assessments

NCLB requires that "all" students should be assessed, and the Federal Register (2007) designated that the assessment based on *modified* academic achievement standards must be aligned with state grade-level content and academic achievement standards. Furthermore, the Federal Register requires that the assessment must be technically adequate. The Federal Register (2007), in the rules and regulations, clarifies the requirement as follows:

> Regardless of whether a State chooses to construct a unique assessment or to adapt its general assessment, any alternate assessment based on modified academic achievement standards must meet the requirements for

> high technical quality set forth in §§ 200. 2(b) and 200. 3(a) (1) (including validity, reliability, accessibility, objectivity, and consistency with nationally recognized professional and technical standards) …(U.S. Department of Education, 2007, p.17750).

Because of the recency of the 2007 requirements, much work needs to be done to determine appropriate methods and approaches. This is particularly true for alternate assessments based on *alternate* academic achievement standards (U.S. Department of Education, 2003) for students with most significant cognitive disabilities, limited recent published research studies on the technical adequacy. Research on technical adequacy issues and score reliability of alternate assessment remains in an initial documentation of conventional psychometric indices including correlation and reliability coefficients and criterion validity (Yovanoff & Tindal, 2007).

The primary reason for the limited research on the technical adequacy of alternate assessments pertains to the challenges associated with obtaining requisite samples of students eligible to take the alternate assessment. The Federal Resister regulations assume students eligible to be assessed based on *modified* academic achievement standards are in regular classrooms and receive grade-level instruction. But in practice, due to their disabilities, they have individualized educational plans (IEPs) designed to meet their unique needs to "access" or consider grade-level curriculum (U. S. Department Education, 2007). Only a small portion of students are eligible for these alternate assessments. According to the regulations, the number of proficient and advanced scores based on the *modified* academic achievement standards does not exceed 2% of all students in the grades tested at the state or local educational agency (LEA) level. This small number of students is designated as the 2% population. The 2% population represents approximately 20 % of students with disabilities (U.S. Department of Education, 2007). The limited number of students within or across grades poses formidable challenges for measurement researchers and state departments of education who are concerned with these issues.

Students with disabilities should be included in accountability systems. This requires accurate assessments for students with disabilities for valid school and student-level educational decisions. However, only limited research has been conducted because of the low prevalence, as well as the lack of perception on the importance of alternate assessment. Considering the low abilities of these students, general education students in lower grade levels may perform similarly to students with disabilities at grade level. If we could determine a procedure for approximating item responses provided by students with disabilities, we will be able to obtain a much easier process for creating measures for this population. One possibility is to test the items with general education students from a lower grade levels.

The purpose of this study was to investigate the sampling procedure for creating mathematics alternate assessment items developed for 2% students. Specifically, this study examined whether psychometric characteristics of mathematics alternate assessment items created for 2% students can be meaningfully estimated with data obtained from general education students in lower grade levels.

The following research questions were addressed in this study.

Research Questions:

1. Are there grade levels at which the *item parameters* estimated from general education students in grades 6-8 are not different from those obtained using the 2% student sample?

(a) Are estimated *item parameters* of the alternate assessment invariant across level of disability when comparing 2 % students in grade 8 with general education students in grade 8?

(b) Are estimated *item parameters* of the alternate assessment invariant across grade level when comparing general education students across grades 6, 7, and 8?

(c) Do estimated *item parameters* depend on an interaction of disability and grade level when comparing 2% students in grade 8 and general education students in grade 6?

(d) Do estimated item parameters depend on an interaction of disability and grade level when comparing 2% students in grade 8 and general education student in grade 7?

2. Are there grade levels at which the estimated *ability* of general education students in grades 6-8 is not different from that of the 2% student sample in grade 8?

CHAPTER II

REVIEW OF THE LITERATURE

The purpose of this chapter is to review the literature related to designing accurate and reliable alternate assessments meeting the requirements of the Federal Register (2007). The review focuses on the recent legislative contexts and practical situations of development and validation of alternate assessments for students with disabilities. This chapter discusses the nature of alternate assessments and the issues in developing alternate assessments. Finally, the present chapter highlights the challenges in evaluating technical adequacy of alternate assessments and presents the contributions of this study.

Alternate Assessments

The NCLB (2001) requires that states must include all students, including students with disabilities, in all state and district-level accountability systems. Alternate assessments are designed for students with disabilities who cannot participate in the regular state assessments, even with testing accommodations. Approximately 10% student with disabilities (i.e., about 1% of the student population) are estimated to participate in the alternate assessment (Thurlow, Elliotte, & Ysseldyke, 2003). Three academic assessment formats including observations, portfolios, and performance assessments were designated by the U. S. Department of Education: 1) teacher observations of students; 2) samples of student work in regular classroom, demonstrating the same content and skills mastery on a computer-scored multiple-choice test covering the same content and skills; or 3) standardized performance tasks such as completion of an assigned task on test day. An alternate assessment must align with the state's content standards, and produce results separately in reading/language arts and mathematics. Also, this assessment should be designed and implemented so that the results indicate adequate year progress (AYP) (U.S. Department of Education, 2003). In order to get an overall picture of performance for all students, the data should be "aggregated" (Thurlow, Elliotte, & Ysseldyke, 2003, p. 78). Participating students in the alternate assessment, regardless of the approaches to measurement such as either observation or testing, must be assessed by the same components that measure general education students and other students with disabilities.

This section discusses the characteristics of alternate assessments based on two sources: NCLB 2003 regulations articulating alternate academic achievement standards for students with the most significant cognitive disabilities and NCLB 2007 regulations articulating modified academic achievement standards for students with persistent learning problems.

Alternate Assessments Based on Alternate Academic Achievement Standards

The U. S. Department of Education released regulations governing the development of alternate assessment for students with the most significant cognitive disabilities identified by existing Individuals with Disabilities Act (IDEA) categories in 2003. The regulations allow states to develop and use alternate achievement standards for students with the most significant cognitive disabilities to determine the adequate yearly progress (AYP) of states, LEAs, and schools. Students should be assessed in the same subjects and test content as all other students, but with a narrower range of content and reduced complexity than the state content standards. The assessment should not focus on testing functional skills or only Individualized Education Programs (IEP) goals.

In order to provide better instruction for all students with disabilities, the U. S. Department of Education expects states to include as many students as possible in academic assessments aligned to regular achievement standards. Even though out-of-level assessments are not alternate assessments, the new guideline of 2003 regulations indicate that out-of-level assessments may be taken into account as alternate assessments to meet the following requirements: 1) aligned with the state's academic content standards; 2) designed to promote access to the general curriculum; and 3) developed to reflect professional judgment of the highest achievement standards possible (U.S. Department of Education, 2003).

Approximately 9% of students with disabilities fit in the most significant cognitive disabilities category. Students counted as proficient and advanced against alternate assessment based on the alternate academic achievement standards cannot exceed 1 %of all students tested in the state or LEA level, by grade and subject. This number represents approximately 10 % of disabled students identified by existing IDEA categories. From this context, the regulations are often referred to as the "1% rule", the eligible students as the “1% population,” and the assessments as the “1% tests” (Burling, 2007a).

Alternate Assessments Based on Modified Academic Achievement Standards

In May 2005, new federal guidelines allowing state’s flexibility in assessing students with disabilities were announced. These guidelines indicate that approximately 2% of students can make progress but may not reach grade-level achievement standards within the year covered by their IEPs. In April, 2007, final regulations articulating alternate assessments based on modified academic achievement standards, also known as the 2% rule, were released (Burling, 2007b). The regulations allow states to develop modified academic achievement standards for a subpopulation of students with disabilities and to adopt and administer assessments based on those standards.

The 2% rule is applied for students for whom the alternate academic achievement standards are too easy, yet the general achievement standards are too difficult. In other words, the 2007 regulations are intended for “gap” students for whom regular state assessments and alternate assessments based on alternate academic achievement standards are inappropriate. The alternate assessment based on this rule must be aligned with grade-level content standards and has “less rigorous expectation of mastery of grade-level academic content standards” (U.S. Department of Education, 2007, p. 17748).

This alternate assessment is targeted for a small number of students with disabilities who can make significant progress, but may not reach grade-level achievement within the time covered by their IEPs. The regulations provide notice that the number of proficient and advanced scores based on the modified achievement standards cannot exceed 2% of all students in the grades tested at the State or LEA level. This number is approximately 20% of disabled students who may have any disability under 13 categories in IDEA (U.S. Department of Education, 2007).

As with the 2003 regulations for assessing students with the most significant cognitive students, the 2007 regulations require technical quality for the alternate assessment-validity, reliability, accessibility, objectivity, and consistency with nationally recognized professional and technical quality. Compared to the 2003 regulations for 1% population, the 2007 regulations do not allow states to use out-of-level assessments for 2% population. The regulations explain that the out-of-level assessments do not cover the same grade-level content as alternate assessments based on modified academic achievement standards.

Lazarus, Thurlow, Christensen, and Cormier (2007) analyzed and reported the characteristics of alternate assessments based on modified achievement standards of six states. Five states including Kansas, Louisiana, North Carolina, North Dakota, and Oklahoma had alternate assessments, and Maryland was developing one. While North Dakota had a portfolio assessment, the other 5 states had a multiple-choice assessment.

The final regulations do not provide clear prescription about which students with disabilities are eligible for alternate assessments based on modified achievement standards. The regulations only indicate that the eligibility will be determined by a student's IEP team. Lazarus et al. (2007) address that the eligibility criteria for the alternate assessment based on modified academic achievement standards differed across 6 states. But students in all 6 states the authors reviewed are required to have IEPs to participate in this assessment. Four out of 6 states included other frequently selected criteria: 1) students are multiple years behind grade level expectations; 2) students were selected for reasons other than students' categorical label; 3) students do not have significant cognitive disabilities; and 4) students were not selected due to student's excessive absences or to social, cultural, environmental, or economic factors. Only three out of 6 states required that students were learning grade-level content, not meeting the criteria of 2007 regulations. The reason, explained by the authors, is that the states which did not require the grade-level content already had assessments or were developing ones for this population before the regulations were enacted. Finally, 2 states required students to have low performance on the state assessment, and 2 states required that the decision should not be based on student's placement setting.

Considering the eligibility issue, Marion (2007) argues that states should describe how many students are eligible to participate in the alternate assessment based on the characteristics of their particular disabilities and the related learning features. In addition to this quantitative information, it is important to present the specific information about the nature of these students' learning, resulting in their failure on the regular assessment: how they learn and how they are taught (Marion, 2007).

Issues in Developing Alternate Assessments for 2% Students

Alternate assessments for 2% students should reflect their unique needs for accurate assessments. Because the requirements in NCLB overtly state that all students must be assessed on grade-level academic standards, it is essential to take into account how to meaningfully assess the skills and knowledge of 2% students (Yovanoff & Tindal, 2007). An alternate assessment should be designed to make the assessment content accessible to the eligible students. This section discusses issues in developing alternate assessments for the 2% population including reducing cognitive complexity as well as test blueprints and test specifications.

Reducing Cognitive Complexity

Designing alternate assessments is a challenge considering the unique cognitive processes of the 2% population. In order to design accurate alternate assessments for this population, the 2007 regulations for modified academic achievement standards provided the following suggestions: (a) replace the most difficult items with simpler items while covering of the state's content standards; (b) modify the same items by removing one of the answer choices in a multiple choice tests; (c) develop a new assessment with flexibility in the presentation of test items using technology such as print, spoken, and pictorial form; and (d) permit students to use dictating responses or use math manipulatives (U.S. Department of Education, 2007).

These strategies focus on reducing cognitive complexity when developing items and implementing tests. However, there is only limited research related to how to reduce cognitive complexity. Even the definition of cognitive complexity is not clear and is controversial. Furthermore, the 2007 regulations, with the above examples, mentioned that modified academic achievement standards, related to alternate assessments, may be "less difficult than grade-level academic achievement standards" for regular assessments (U. S. Department of Education, 2007, p. 17750).

Despite the lack of clarity, there exists some information relevant to cognitive complexity. Cognitive complexity has been addressed primarily in the psychology field in relation to cognitive process and reasoning, and in general cognitive complexity means levels of thinking (McDaniel & Lawrence, 1990). McDaniel and Lawrence (1990) asked students to write interpretations of two situations after providing video and written materials. Through the students' responses, the authors address 5 levels of cognitive complexity: 1) unilateral descriptions, 2) simplistic alternatives, 3) emergent complexity, 4) broad interpretations, and 5) integrated analysis. The authors explain that students in level 1 just simplified the situation, simply paraphrased, restated or repeated information, while students in level 2 were able to recognize simple and obvious conflicts, and address alternatives. Students in level 3 presented more than one perspective and started establishing complexity by supporting their position through comparison and simple causal statement. Level 4 students could use and integrate many ideas for the interpretation. Finally, students in level 5 were able to construct conceptual frameworks and predict results. From their studies, the authors argue that the cognitive complexity level was related to developmental levels (i.e., grade levels), at least through the high school years.

In the education field, cognitive complexity has been studied through content analysis to achieve content evidence validity for the alignments of assessments to content standards. Webb (1999) identifies four criteria to explore the alignment of assessments to standards: Depth of Knowledge (DOK), Categorical Concurrence, Range of Knowledge, and Balance of Representation. Webb stated that the DOK levels are associated with cognitive complexity levels and classified it with 4 levels: Level 1 refers to recall, level 2 refers to skill or concept, level 3 refers to strategic thinking, and level 4 refers to extended thinking (Webb, 1999; 2002; 2007). Webb (2007) agreed with McDaniel and Lawrence (1990) that DOK level, to some extent, depended on grade levels and the knowledge and performance expectations of a traditional student working at grade level. In addition, the "sophistication of the passages" used in testing could be a variable of increasing complexity across grades (Webb, 2007, p. 22). For example, as sophistication of passage increases, student may find test problems more difficult even though the main concept of the passage is constant. However, if the complicated problem demands more inferences or paraphrasing, the author addresses that DOK level may be increased across grades. The author concludes that there are no rigorous guidelines for acceptable progression in content complexity from one grade to the next grade. Finally, Webb (2007) argues that because DOK levels describe the complexity of content through a content analysis, DOK levels are "related to" cognitive levels, but this does not mean that the levels "correspond to" cognitive levels (p. 24). While Webb articulated DOK with 4 levels, Porter (2002) considers cognitive complexity as "level of depth and specificity" of contents with 5 categories (p. 3), using content-by-cognitive level matrix. The author defines the cognitive demand dimension with five categories: 1) memorize facts, definitions, and formulas; 2) perform procedure/ solve routine problems (e.g., doing computations, taking measurements); 3) communicate understanding (e.g., communicating mathematical ideas); 4) solve nonroutine problems (e.g., applying mathematics in context outside of mathematics); and 5) conjecture/generalize/ prove (e.g., inferring from data and predicting).

Bloom's (1956) original Taxonomy includes 6 major categories in the cognitive domain—Knowledge, Comprehension, Allocation, Analysis, Synthesis, and Evaluation. The domains were arranged by the level of cognitive complexity: from simple to complex and from concrete to abstract with a cumulative hierarchy. In other words, a higher level of category (e.g., Synthesis) requires more cognitive complexity than a lower level (e.g., Knowledge). In addition, the mastery of each simpler category was required to move forward to mastery of the next more complex one. Krathwohl (2002) revised the Bloom's Taxonomy to two-dimensional framework: Knowledge and Cognitive Process. Knowledge in Krathwohl's framework is comparable to the subcategories of Knowledge in Bloom's Taxonomy. Cognitive process in Krathwohl's framework is parallel to the six categories of the Bloom's Taxonomy. Three of the 6 categories were renamed: Knowledge to Remember, Comprehension to Understand, and Synthesis to Create. The last 3 was changed to the verb forms: application to apply, analysis to analyze, and evaluation to evaluate. These new 6 categories were also ordered in a hierarchy, while not so rigorous as in the original Bloom's Taxonomy. Finally, the structure of the Knowledge dimension includes 1) factual knowledge, 2) conceptual knowledge, 3) procedural knowledge, and 4) metacognitive knowledge. The structure of the cognitive process dimension has 1) remember, 2) understand, 3) apply, 4) analyze, 5) evaluate, and 6) create.

Cognitive complexity appears to involve cognitive processing. Gorin (2006) suggests a specific item level framework of cognitive processing—item difficulty modeling (IDM). The author addresses a specific item level of cognitive model providing valuable information in students' performance. IDM consists of a series of cognitive processes or skills arranged by the sequence of item processing. According to the author, item difficulty may be a function of two general processes: inferring rules and applying them. In addition, the cognitive complexity of these processes could result from the number of rules and the complexity of the rules. To develop a model of cognitive complexity for an IDM reading comprehension example, the author used two processes: 1) Text representation and 2) Response decision. Cognitive complexity and therefore student performance may be affected by these two processes. Text representation includes Encoding (e.g., vocabulary level of the passage and propositional density of the passage) and Coherence Process (e.g., percent content words and passage length). Response decision consists of Encoding and Coherence processes (e.g., vocabulary level of the question, the correct response, and the distractors), Text Mapping (e.g., reasoning level of the question, the correct response, and the distractors), and Evaluate Truth Status (e.g., confirmation level of the correct response and number of distractors) (Gorin & Embretson, 2006). Gorin (2006) addresses this kind of item-specific cognitive model may be helpful to detect construct-irrelevant variance and eventually to investigate construct validity at the item level.

While there is not a fully accepted definition of cognitive complexity, there are some similarities in the definitions of cognitive complexity. First, the authors analyzed the cognitive processes when students respond to test items and tried to figure out how cognitive complexities affect measuring students' performance. Second, the levels of cognitive complexity the authors suggested are ordered from simple to complex or easy to difficulty, and related to developmental levels (i.e., grade levels). Lastly, and most unfortunately, is that the issue of cognitive complexity has not been applied to working with 2% students. These essential facets of cognitive complexity will guide proposed procedures in developing accurate assessments with this unique population.

Implications of Cognitive Complexity Research on 2% Alternate Assessments

Taking a close look at these studies provides some specific guidelines for designing alternate assessments for 2% students. Webb (1999) addresses that item format may be a confounding factor of DOK level. The author illustrated the item format issue with a graph interpretation item by two different versions: open response version and fixed response version (i.e., multiple-choice test). The multiple-choice item could be a level 2 and the open-response item a level 3 because the multiple-choice item can be solved by removing the other distractors. This implies that 2% students may benefit from changing the item format to multiple-choice instead of open response. Additionally, Webb (2007) argues that the sophistication of the passages could be one variable of increased cognitive complexity. Simplifying passages while retaining the main idea or concept may be one way to design alternate assessments. Furthermore, the IDM addressed by Gorin (2006) drives us to consider the complexity of text representation and response decision processes at an item level. As with assessments for other general education students, each subcomponent of IDM (e.g., vocabulary level and length of passage, reasoning level of the question, confirmation level of the correct response, and the number of distractors etc.) must also be considered for assessments for the 2% students.

Practical examples for reducing cognitive complexity to design alternate assessments for the 2% students can be found in Lazarus et al. (2007)'s analysis of 6 states. The authors reported that four states reduced the number of answer choices of multiple-choice items from four possibilities to three possibilities. Also, 4 states had simplified language and fewer items than regular state assessments. Furthermore, shorter reading passages were applied to three states. Finally, there was a state using segmented reading passages directly followed by the questions, and one having all items with 1 or 2 levels of DOK.

Test Specifications and Test Blueprints

The content is a critical factor in developing assessments. A test consists of items in the domains reflecting the construct intended to be measured. The items in a test should cover the full range of intended domains with a sufficient number of items. The construct may be underrepresented if the number of items is too few to measure student knowledge and skills or the test does not cover all required content (Messick, 1989). To avoid this risk and confirm content coverage, test developers document test specifications and test blueprints. Test specifications are created to establish the guidelines about the test content and test items. Based on the guidelines, a test blueprint provides item format and the number of questions to be written in each content category (Oregon Department of Education, 2007).

In addition to the information about test content and test items, test specifications include the amount of testing time, directions for test takers, test administration instruction and scoring procedures (AERA, APA, & NCME, 1999). For example, the *Oregon Mathematics Test Specifications* includes the description about the applied item format (i.e., multiple-choice format), the conversion of the raw score to a scale score (i.e., RIT score), and the description of the content. In addition to this information, test specifications for alternate assessments may need to describe who to test and how to identify the students (U.S. Office of Special Education Programs, 2006).

Test blueprints include information about individual tests including test content, the number of test items by each content category, and the formats of those items such as short-answer, multiple-choice or extended-response. Test blueprints are delivered to educators via test manuals (U.S. Office of Special Education Programs, 2006). For example, The Oregon Department of Education indicated that the blueprint for each mathematics test must include the following components: 1) Score Reporting Categories (SRC); 2) the cognitive demand and difficulty level of items; 3) the number and percentages of test items from each SRC on the test; and 4) the total number and percentages of operational and field test items. Score reporting categories include calculations and estimations, measurement, statistics and probability, algebraic relationships, and geometry. There are three different versions of test formats—paper/pencil administration and electronic administration including short online and long online form. Table 1 presents the grade 8 mathematics test blueprints.

Table 1

Mathematics Test Blueprint—Content Coverage and Weighing of Grade 8

Score Reporting Categories	Number of KS items Long Test	Number of KS items Short Test	Number of KS items P/P Test	% of questions assessed per test[†]	Online Test Pool size
Calculations and Estimations	6-9	4-7	9	15% (12-18)	50-100
Measurement	6-9	4-7	9	15% (12-18)	50-100
Statistics and Probability	9-11	7-10	12	20% (18-22)	50-100
Algebraic Relationships	15-18	9-12	18	30% (30-36)	50-100
Geometry	9-11	6-9	12	20% (18-22)	50-100
Operational Item Total	50	35	60		
Field Test Item Total	5	5	NA		
Total Items on Test	55	40	60	100%	

[†] ( ) display range of distribution allowed to ensure maximum precision of students' total scale score. KS = Knowledge and Skills.

The mathematics tests are designed considering a range of DOK items, and a range of difficulty items are included. Three DOK levels are applied to Oregon's multiple-choice test items: 1) Recall (recall a fact, information or procedure); 2)

Skill/Concept (use a skill or a concept including two or more steps thinking); and 3) Strategic Thinking (use a reason, develop a plan or use a sequence of steps). All raw scores are converted to scaled scores based on the number of questions answered correctly compared to the total number of questions on the form. This scaled score is called a Rasch unit or RIT score, and may be considered as the difficulty of the items. In the paper/pencil test, items of medium or easy difficulty appeared first by followed by more difficult items. In online tests, students are given mean RIT level items and according to students' correct and incorrect responses, the following items are selected (Oregon Department of Education, 2007).

Test Specifications and Test Blueprints for 2% Alternate Assessments

What is the content for the alternate assessment for the 2% population? The 2007 regulations require alternate assessments to have the "same grade-level content as the regular assessment" (p. 17750) with less difficulty than the state assessment. Regardless of which assessments students would take to be evaluated, the student should be taught and assessed in grade-level content. In other words, the purpose of defining content and achievement domains should be for instruction as well as assessment. The instructional needs of participating students should be the first priority when defining the domain. In this context, test blueprints for both regular state assessments and alternate assessments would provide valuable information to define the domain (Marion, 2007).

Test blueprints for alternate assessments may include similar information in the regular state assessment because the content for both assessments should be same. For example, the Oklahoma State Department of Education (2006) designated that the blueprints for the alternate assessment based on modified academic achievement standards use the same percentage of items per standards as the original blueprints for the regular state assessment. The number of items for a mathematics alternate assessment for the 2% students in 8th grade is 33, while the number for the regular state assessment is 45. The California Department of Education is developing alternate assessments based on modified academic achievement standards, titled the California Modified Assessment (CMA). The mathematics CMA blueprints were developed for grades three through seven. The blueprints consist of 48 or 54 items depending on the grade level, while the California Standards Mathematics Tests contain 65 items. The blueprints include the number of items for each strand and similar percentage of items per strand to the California Standards Tests (California Department of Education, 2008). In the case of including fewer items in alternate assessments than in regular assessments, the items should be representative of the intended domain to avoid the risk of construct underrepresentation (U.S. Office of Special Education Programs, 2006).

Technical Adequacy of Alternate Assessments

The final regulations require that alternate assessments based on *modified* academic achievement standards demonstrate the same technical quality as other regular state assessments. In practice, however, there have been several challenges to developing accurate, reliable, and accessible assessments for this population (Thurlow, Elliot, & Ysseldyke, 2003). This section discusses the importance of technical adequacy of alternate assessments and the challenges in evaluation of the technical adequacy of alternate assessments. Additionally, the differential item functioning (DIF) is discussed as a consideration to solve the sampling issue, one of the challenges.

Challenges in Evaluation of the Technical Adequacy of Alternate Assessments

"Use of appropriate and technically defensible assessments" is a critical consideration in developing assessments (Rabinowitz & Sato, 2006, p. 8). Once alternate assessments are created for 2% population, the state must demonstrate the technical adequacy (i.e., reliability, validity, accessibility, objectivity, and so on) of the alternate assessments. Technically adequate alternate assessments allow states to fully and equitably include these students in their accountability systems. Little research on the technical adequacy of the assessments has been done because the assessment of 2% population is a relatively new area of work (Rabinowitz & Sato, 2006).

Targeted population. The biggest difference in evaluating technical adequacy between regular state assessments and alternate assessments must be the targeted population. Due to the different population, the appropriate procedures and criteria for determining the technical adequacy of 2% assessments may not completely comply with the procedures and criteria for the regular state assessments. Sato et al. (2007) argue that the technical adequacy of assessments for special population must be reviewed with respect to the particular population. To guarantee that the assessment is accurate and reliable for the targeted population, the population should be well defined. This effort is essential to avoid the risk of misunderstanding of the population between test developers and users (i.e., educators, students, and parents). Without this consideration, the valid interpretation of student scores may not be guaranteed (Rabinowitz & Sato, 2006). Ultimately, this compromises the fundamental goal of assessments, i.e., designing and providing appropriate instruction.

Sampling issue. When addressing the importance of the target population, Rabinowitz and Sato (2006) also present concerns about the difficulty in conducting the reliability and validity studies for students with disabilities. A relatively small number of the population may cause sampling issues, resulting in high costs and difficulty in implementation of research (Rabinowitz & Sato, 2006). Generally speaking, the larger the sample size, the more likely the research results will be generalizable because the sample will more likely reflect the population. Furthermore, larger appropriate sample sizes are necessary for parameter estimation with acceptable standard error. However, in practice, almost 10% students are considered students with disabilities. Even within those students, there may be several subgroups, including their disability categories, academic or behavioral characteristics, grade levels, and residential location depending on specific research questions. The Federal Register (2007) assumes that the eligible students for alternate assessments based on *modified* academic achievement standards comprise approximately 2% of the total number of students and 20% of the students with disabilities. Working with a subset of an already small group poses a significant challenge for conducting research on adequacy for alternate assessments. Due to this challenge most researchers have resorted to using out -of-level testing procedures — testing students a level below their grade level (Thurlow, Bielinski, Minnema, & Scott, 2002). However, the 2007 regulations do not allow states to use out-of- level assessments since they do not meet the "same" grade-level content requirement.

To solve the sampling issue, considering the low abilities of these students, is it simply assumed that general education students in lower grade levels may perform similarly to students with disabilities at grade level? Without any evidence, is it possible to use lower grade level of general education students' performance on the grade-level test to validate measures created for special education population? Angoff (1993) notes that age difference may produce different intellectual behaviors even with the same mental age. For example, an 8th grade student and a 6th grade student may perform differently even with the same mental age of a 6th grader. The author suggests that differential item functioning (DIF) techniques, one application of the Item Response Theory (IRT), can be used to find out what the nature of these differences is and where these differences occur.

Differential Item Functioning (DIF)

Overview of Item Response Theory (IRT). IRT has been applied routinely in the development of standardized tests. In classical testing theory, the same item could be considered as an easy or as a difficult one depending on the average ability of the examinees sample. If the sample is highly skilled, an item may appear easy, while that same item will be difficult for another low ability sample. To the contrary, IRT provides sample invariant item statistics, e.g., item difficulty. That is, there is no difference between item parameter estimates even though the data analyzed were obtained from two groups with different average abilities.

One of the most popular IRT models is the unidimensional one-parameter logistic (1 PL) model, also known as the Rasch model. In the 1 PL model, there is one item parameter (β) and one person ability parameter (θ). The item parameter is a calibration of item difficulty, and the person parameter is a calibration of person ability. Two item characteristic curves (ICCs) are illustrated in Figure 1, showing a relation between an item difficulty, person ability, and the probability of responding correctly to the item P(X=1). First, P(X=1) is a monotonically increasing function of ability, i.e., as ability increases so does the probability of a correct response. Second, as item difficulty increases, more ability is necessary for a correct response. For the 1PL model, item difficulty (β) is that point on the difficulty scale at which the probability of a correct response is 0.50. In the Figure 1, item 2 is more difficult than item 1.

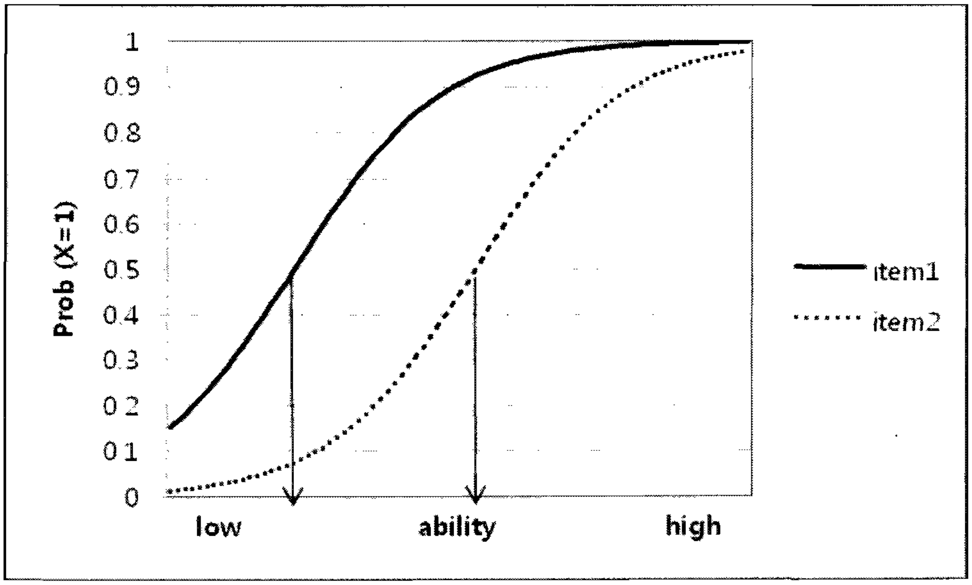


Figure 1. Item Characteristic Curves (ICCs)

To estimate item and person parameters for IRT models, two key assumptions are required: (a) local independence and (b) appropriate dimensionality. These two assumptions are very closely related. Local independence implies that the response to any item should not be affected by any other item if trait level is controlled. In other words, even though the items are highly correlated in the sample, if trait level is controlled, there will be no relationship between items. Item intercorrelation should emerge only as trait level varies. Local independence is often interpreted as unidimensionality. Appropriate dimensionality is the inclusion of the necessary number of latent traits essential for properly modeling item response. Most measures are developed to as though they are unidimensional, i.e., response to items is a function one skill only. If a response requires more than one skill, then the appropriate dimensionality of the item response model must include parameters for multiple skills, e.g., multidimensional. How well IRT models fit the data is affected by appropriate dimensionality. While researchers continue to study multidimensional IRT modeling, most IRT models assume a single latent-trait dimension (i.e., unidimensionality) (Embretson & Reise, 2000).

Overview of Differential Item Functioning (DIF). Another measurement consideration often related to measurement dimensionality is DIF. DIF models are used to investigate measurement bias, testing hypotheses regarding sample invariance, i.e., item parameters are constant for different respondent populations. For instance, referring to Figure 1, DIF analyses test the hypothesis that the item parameter β does not change as respondent characteristics change, e.g., male, female. Persons with the same ability should have the same probability of a correct response. If persons with same ability have different probabilities, then there may be other measurement dimensions operating in the response process. Angoff (1999) illustrates that DIF was initially applied to figure out the reason for the great gap in performance between Black and Hispanic and White students on tests of cognitive ability. The assumption behind the study was that students in the minority cultures were not familiar with the test contents. So DIF was used to detect any biased items against the minority students and eliminate the items. An item is said to be biased when students of equal ability from different groups do not have equal probabilities of responding the item correctly (Angoff, 1999).

According to the *Standards for Educational and Psychological Testing* (1999), DIF occurs when students of equal ability differ on average in their response to a particular item, depending on their group membership. Also, the authors of the *Standards for Educational and Psychological Testing* consider DIF as construct-irrelevant variance that differentially affect the test scores for students in specific groups. When responding to an item, students with equal ability should have the same probability of providing a correct answer, irrespective of their group membership, e.g., students with a disability and students without a disability. Importantly, the existence of group mean difference and standard deviation do not indicate the presence of DIF (Thissen, Steinberg, & Gerrard, 1986). Item and test characteristics are assumed to be invariant, stable, across identity groups, though the groups may be quite different with respect to mean ability. A test item is referred to as showing DIF if the item-response curve (IRC) is not the same for two groups (see Figure 2 and Figure 3) typically referred to as the reference and focal groups (Embretson & Reise, 2000). However, the existence of DIF in an item does not necessarily mean that the item is biased and unfair to one of the tested groups.

Even though the DIF interpretation may differ on research purpose and context, there is general agreement about when DIF is to be considered and interpreted. Items showing DIF for two or more groups are considered to violate the unidimensionality assumption, one of the fundamental assumptions of IRT. Large DIF would be a sign that the item is measuring an additional, unintended construct of the test (i.e., construct-irrelevant variance). This implies that the item is not unidimensional for at least one of the tested groups, or does not measure the same construct (dimension) in tested groups. As noted above, appropriate dimensionality has not been achieved in the modeling process. In this case, the construct measured by the item may be different from one group to another group, rendering test inferences invalid if based on the test scores including the item. So the test can be said to be biased or unfair (Angoff, 1999). That is, a DIF item is taken into account as biased when the construct-irrelevant component placing students in one group at a disadvantage is identified (Gierl, 2005). Recalling the mathematics item presented previously, the presence of construct-irrelevant variance is likely to result in DIF, i.e., the presence of measured differences among two students/groups, though the actual abilities may be equal.

There exist two different kinds of DIF—uniform DIF and non-uniform DIF. Mellenberg (1982) defines these two concepts by interaction between ability level and group membership. Uniform DIF is said to present when there is no interaction between ability level and group membership with respect to probability of a specific response. In other words, the probability of answering the item correctly for one group is uniformly larger (or smaller) than the other group over all levels of ability. In this case, the item characteristics curves (ICC) for both groups do not cross, while there is uniform difference between the ICCs (see Figure 2).

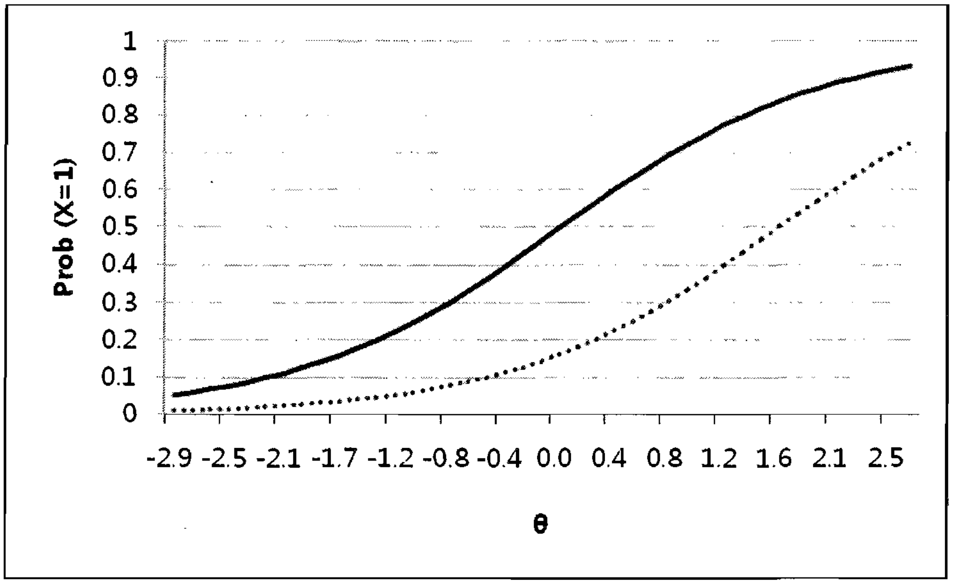


Figure 2. ICCs with Uniform DIF.

In contrast, non-uniform DIF exists when there is interaction between ability level and group membership (see Figure 3). The difference in the probabilities of a correct answer for the two groups depends on the ability levels. In this case, the ICCs cross over one another (Swaminathan & Rogers, 1990). Better understanding of the nature of DIF is as important as identifying items that are functioning differently.

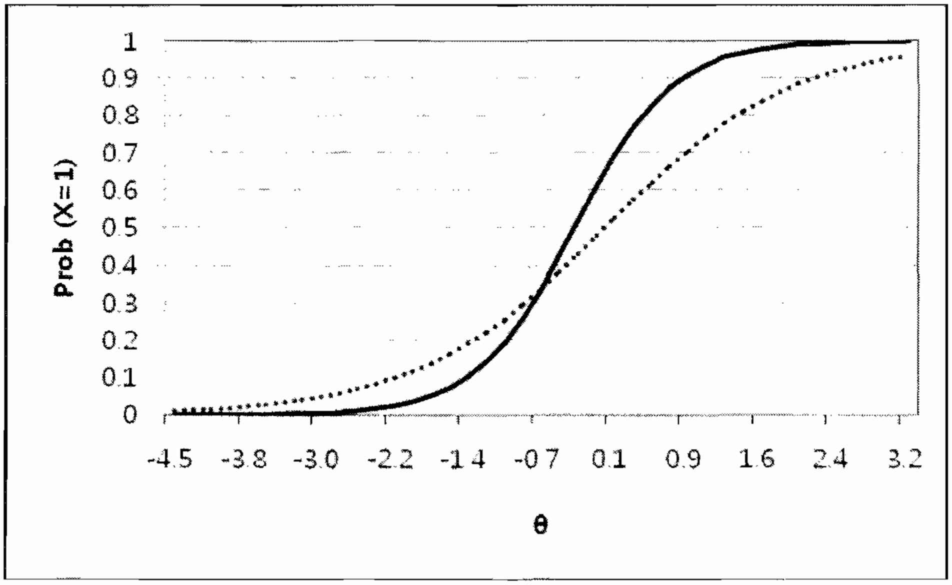


Figure 3. ICCs with Non-uniform DIF.

There are diverse approaches for DIF detection including Mantel-Haenszel (MH) and logistic regression (LR). Crane, Belle, and Larson (2004) compared MH and LR approaches and conclude that MH approaches are more appropriate for the uniform DIF analysis, while LR techniques are useful for both uniform and non-uniform DIF detection. MH is the most widely used technique for DIF detection (Clauser & Mazor, 1998; Holland & Thayer, 1998). Zwick and Ercikan (1989) cited Holland and Thayer (1988)'s description about MH procedure. Data are obtained from the reference group R and the focal group F. According to the description, given the sample sizes (n_{Rk} and n_{Fk}) and the probability of answering the item correctly (p_{Rk} and p_{Fk}), A_k and C_k are independent binomial random variables with parameters of (n_{Rk} , p_{Rk}) and (n_{Fk} , p_{Fk}). The null hypothesis for MH is that there is no relation between group membership and the test performance after controlling for examinee's ability. Mathematically, the null hypothesis for the MH procedure is expressed by

$$H_0 = \frac{p_{Rk} / q_{Rk}}{p_{Fk} / q_{Fk}} = 1, \quad \text{k=1, 2,.....K.}$$

And the alternate hypothesis is

$$H_1 = \frac{p_{Rk} / q_{Rk}}{p_{Fk} / q_{Fk}} = \omega, \quad \omega \neq 1.$$

Zwick and Ercikan (1989) describes that the parameter ω refers to the common odds ratio for the K 2 x 2 tables. Finally the MH chi-square statistic is defined by

$$\chi^2_{MH} = \frac{\left(\left|\sum_k A_k - \sum_k E(A_k)\right| - \frac{1}{2}\right)}{\sum_k Var(A_k)}$$

where $E(A_k) = n_{Rk} m_{1k} / T_k$ and $Var(A_k) = \dfrac{n_{Rk} n_{Fk} m_{1k} m_{0k}}{T_k^2 (T_k - 1)}$.

The MH procedures also provide the estimator of ω:

$$\widehat{\omega}_{MH} = \frac{\sum A_k D_k / T_k}{\sum B_k C_k / T_k}.$$

The $\widehat{\omega}_{MH}$ means the ratio of the odds that a reference group examinee will answer correctly compared to the odds for a matched focal group examinee. Educational Testing Service (ETS) provided an index of differential item performance:

MH D-DIF= Δ_{MH} = -2.35ln ($\widehat{\omega}_{MH}$).

Using this formula, Zwick and Ercikan (1989) suggest the following interpretation guidelines to evaluate the DIF effect size:

1. Negligible or A level DIF: MH test is not statistically significant or $|\Delta_{MH}| < 1$
2. Moderate or B level DIF: MH test is statistically significant and $1 \leq |\Delta_{MH}| < 1.5$
3. Large or C level DIF: MH test is statistically significant and $|\Delta_{MH}| \geq 1.5$

Zheng, Gierl, and Gui (2007) illustrate another approach for DIF detection, the logistic regression (LR) procedure. According to the authors the equation for DIF detection is expressed as

$$P(u = 1|\theta, g) = \frac{e^{f(\theta,g)}}{1 + e^{f(\theta,g)}}.$$

$P(u = 1|\theta, g)$ is the conditional probability of answering correctly given the observed ability (θ) and the group membership (g). In addition, the function $f(\theta, g)$refers to the linear combination of the predictor variablesθ, g, and the interaction θg. Givenτ_0 and $\tau_1$represent the intercept and weights for the ability, there are three models in LR procedure:

1. Model 1: $f(\theta, g) = \tau_0 + \tau_1\theta$,
2. Model 2: $f(\theta, g) = \tau_0 + \tau_1\theta + \tau_2$
3. Model 3: $f(\theta, g) = \tau_0 + \tau_1\theta + \tau_2 + \tau_3\theta g$.

Zheng, Gierl, and Gui (2007) describe 3 different steps for DIF detection with the above three models. While model 1 serves as the baseline model in step 1, in step 2 the occurrence of uniform DIF is tested by investigating the improvement in chi-square model fit related to addition of group membership (g) against the baseline (i.e., Model 1-

Model 2). In step 3, the occurrence of non-uniform DIF is tested by investigating the improvement in chi-square model fit related to addition of group membership (g) and the interaction between test score and group membership (θg) against model 2 (i.e., Model 2-Model 3). Zheng, Gierl, and Gui (2007) introduced the Jodoin and Gierl (2001)'s guidelines for effect size measure for uniform DIF detection with LR procedure, called $R^2\Delta - U$:

$$R^2\Delta - U = R_2{}^2 - R_1^2$$

where $R_2{}^2$ and $R_1{}^2$ are the sums of the products of the standardized regression coefficient for each explanatory variable and the correlation between the response and each explanatory variable of the model 2 and model 1. The guidelines to evaluate DIF effect size suggested by the authors as follows:

1. Negligible or A level DIF: $R^2\Delta - U$ <0.035
2. Moderate or B level DIF: Null hypothesis is rejected and

   $0.035 \le R^2\Delta - U < 0.07\,0$

3. Large or C level DIF: Null hypothesis is rejected and $R^2\Delta - U > 0.07\,0$.

DIF has not been used often in research related to the special education field because it is sensitive to sample size. However, Angoff (1999) notes that DIF techniques may help answer relevant research questions. Why do students with special needs perform differently from their peers even with the same mental age? Are there any particular cognitive dimensions for students with disabilities? The author explains that these questions can be transferred into the context of test performance: how are these differences identified in the student's responses to the item? For multiple choice items, is performance related to the correct answer choice or the distractors? If these differences do not occur in both correct answer choice and distractors, why do these differences exist? Angoff (1999) suggests the potential for DIF analyses in context of these types of questions.

Contributions of the Present Study

The recent 2007 regulations allow states to adjust statewide testing programs to 2% students. These students did not benefit from either state regular assessments or 1% alternate assessments. The sampling issue has been a significant challenge in developing alternate assessments for 2% students. The present study investigated the sampling procedure for developing mathematics alternate assessment items designed for 2% students in grade 8. The results of this study may provide new procedures for developing alternate assessments for 2% students that can expedite test development across content areas and grade levels.

CHAPTER III

METHODS

The purpose of this study was to examine the sampling procedure for developing mathematics alternate assessment items designed for 2% students. Specifically, the present study investigated whether psychometric characteristics of mathematics items for 2% students in grade 8 could be meaningfully estimated with data obtained from general education students in lower grade levels. This chapter discusses the procedures of collecting participants and the verification of participants for this study. Next, a description of the measures used in the present study is provided. Last, the procedures of data collection and data analysis are developed and discussed in relation to the research questions.

Participants

The presents study was conducted during the spring of 2008 via online assessment in Oregon. First, six general education teachers signed up for this study via computer with 4 of 6 teachers enrolling their students (N=347). The number of enrolled students by each teacher ranged from 12 to 168. Two hundred thirty students had completed their test whereas 241 students had started a test. Finally a total of 235 general education students from grade 6, 7, and 8 were verified for this study. Second, in order to collect data from students with disabilities, 11 teachers were recruited, of which 10 teachers enrolled their students. The number of registered students by each teacher varied from 2 to 11. A total of 51 students with disabilities including grades 6, 7, and 8 were signed up, of whom 47 students had started a test and all 47 students had completed their test. Only 39 students with disabilities out of 43 in grade 8 were confirmed for this study because 4 students had not started a test even though their teachers completed the Teacher Perception Survey for them. In sum, a total of 235 general education students in grades 6, 7, and 8 were verified. Thirty nine students with disabilities in grade 8 were confirmed for this study. Table 2 presents the specific number of students by grade and disability levels for this study.

Table 2

Number of Students by Grade and Disability Levels

	Grade		
	6	7	8
General Education Students	113	52	70
Students with Disabilities	-	-	39

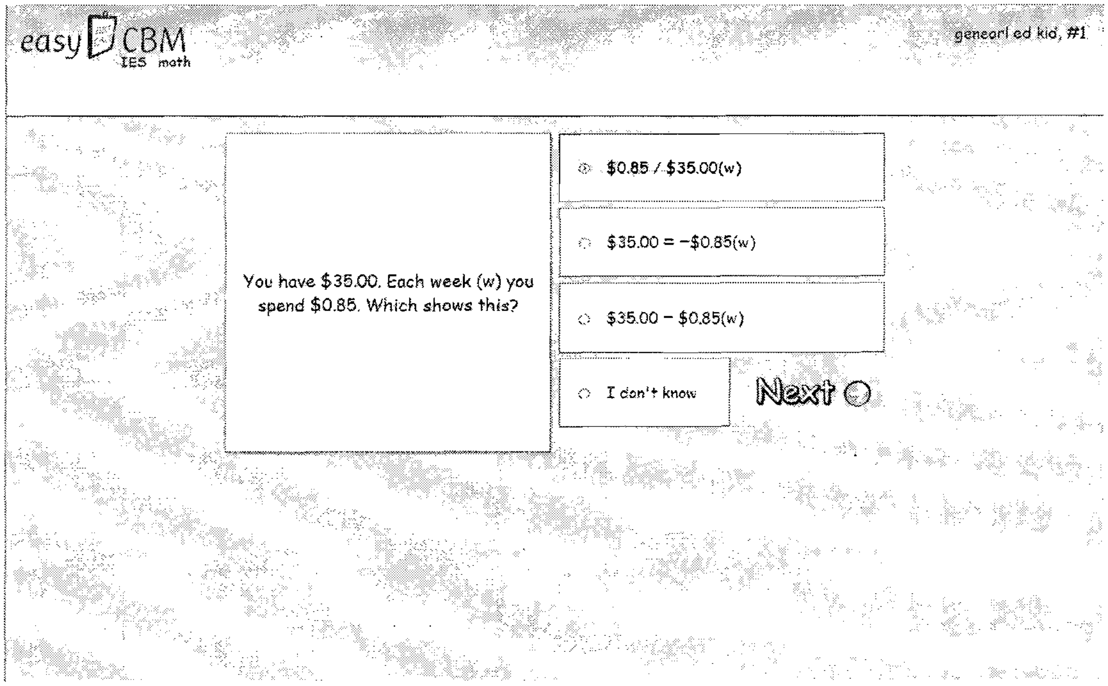
Measures

Three different types of items were used in this study: 1%, 2%, and standard items. All of the items were matched in terms of content standards (i.e., objectives) across the three test types. The Student Performance Test included 2% items and standard items, while the Teacher Perception Survey consisted of 1%, 2%, and standard items.

Student Performance Test

A total of 20 mathematics items were administered to all participants in a single testing: 10 2% items and 10 standard items. 2% items were written to address a draft of the *Oregon Mathematics Content Standards*. These standards were based on *the Curriculum Focal Points for Pre-Kindergarten through Grade 8: A Quest for Coherence in September 2006.* Each item is linked to a specific objective of the 8th grade level's draft of the *Oregon Mathematics Content Standards* (see Appendix A for item alignment), but many items could link to more than one objective and/or more than one grade level. Three domains are included in *Oregon Mathematics Content Standards*: 1) Algebra; 2) Geometry and Measurement; and 3) Data Analysis, Number and Operations, and Algebra. All 2% items were written in English, using simple vocabulary and short declarative sentences. The items were developed with reduced cognitive complexity while retaining the 8th grade level content standards (see Appendix B for 2% item-writing guideline). In addition to these 10 items, performance on10 items that were released from the 8th grade *Oregon Department of Education Sample Test* were investigated.

The test was delivered to students and scored via computer. For this study, 20 items were arranged with 2% items as the odd number (i.e., item 1, item 3, item 5, etc.) and standard items as the even number (i.e., item 2, item 4, item 6, etc.). Items were formatted with the question on the left of the screen and the answer choices listed vertically on the right. Figure 4 illustrates this figuration. Each item with the 4 option multiple choice response format, including the option, “I don’t know,” were presented on the screen. The “I don’t know” option was included so that participating students felt less failure when responding to items. In addition, the option was intended to control a guessing factor. Students were able to select the answer choice by clicking anywhere within the box containing an answer choice. When the computer delivered the items, it randomly sorted the answer options, capturing which of the selections the students choose. Students were allowed to change their answer choices at any time prior to submitting their responses by clicking the “NEXT” arrow. Students were permitted to use scratch paper and a pencil, while calculators were not allowed. After answering a question, students were not able to go back.



You have $35.00. Each week (w) you spend $0.85. Which shows this?

- ○ $0.85 / $35.00(w)
- ○ $35.00 = −$0.85(w)
- ○ $35.00 − $0.85(w)
- ○ I don't know

Figure 4. Interface Design for Test Items.

Teacher Perception Survey

The Teacher Perception Survey consisted of 15 sets of 1%, 2%, and standard items. Two percent and standard items for the Teacher Perception Survey had the same characteristics described above as the items for the Student Performance Test. One percent items used for the Teacher Perception Survey came from the *2007-2008 Oregon Extended Assessments*. In the *2007-2008 Oregon Extended Assessments*, the *Middle School Assessment* were used for students in grades 6, 7, and 8, while the *Elementary Assessment* were applied to students in grades 3, 4, and 5. In order to conduct the Teacher Perception Survey for students in grade 8, 15 items were released from the *Middle School Assessment*.

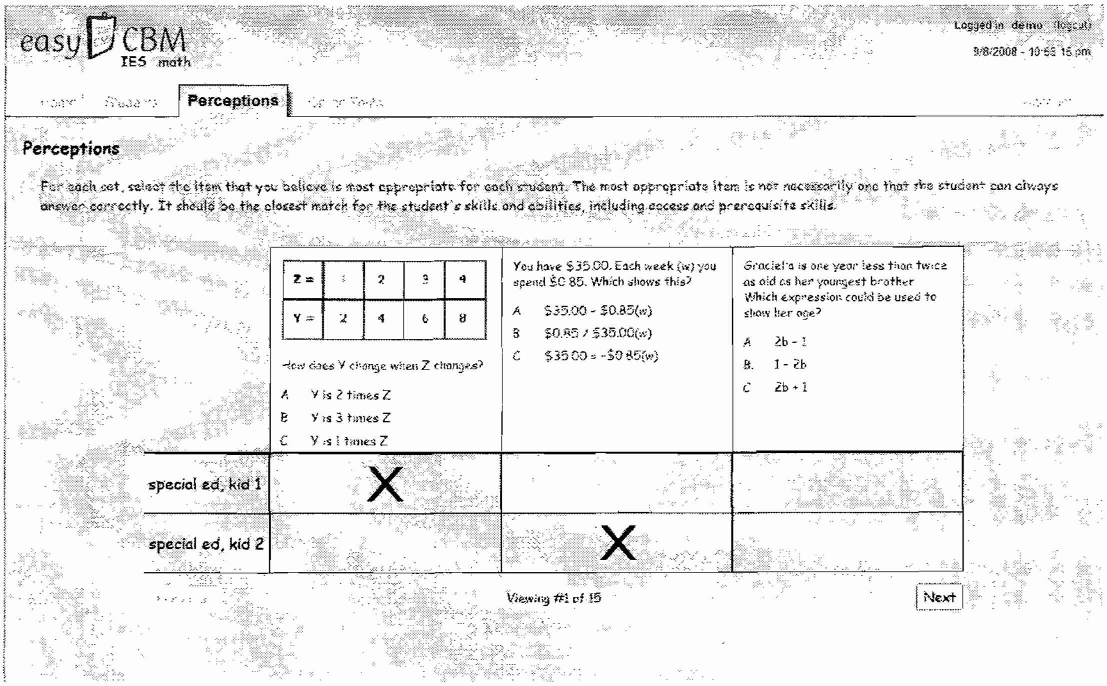
Procedures

Data Collection

Data were collected from both students with disabilities and general education students. First, identification of the 8th grade 2% students was based on two data sources: (1) Teacher Perception Survey and (2) Student Performance Test.

Teacher perception survey. Teachers were shown 15 sets of items including 1%, 2%, and standard items. Three out of 15 sets included only 1% and standard items. The draft of the *Oregon Mathematics Content Standards* for the 2% items are the most recent standards, and the state's 1% and standard items were written to an older set of content standards. Some of the 1% and standard content standards did not appear in the 2% test specifications. In these cases (i.e., sets 2, 8, and 15), the set did not include the 2% item.

For each set, the teachers indicated which item (1%, 2%, or standard item) was the most appropriate for the students considering the students' skills and abilities including access and prerequisite skills. Figure 5 illustrates this configuration. Three items for each set were randomly ordered so that teachers were not able to notice which item is 1%, 2% or standard item. Teachers were able to select the item by clicking anywhere within the box below the item in the row of the student's name. Teachers were allowed to change their opinion at any time prior to submitting their responses by clicking "NEXT". Even after selecting a specific item for each student, teachers were allowed to revisit the previous set of items whenever necessary. Teachers were required to respond to all enrolled students to move forward to the next set of items.



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

	Z = | 1 | 2 | 3 | 4 Y = | 2 | 4 | 6 | 8 How does Y change when Z changes? A Y is 2 times Z B Y is 3 times Z C Y is 1 times Z	You have $35.00. Each week (w) you spend $0.85. Which shows this? A $35.00 - $0.85(w) B $0.85 / $35.00(w) C $35.00 = -$0.85(w)	Graciela is one year less than twice as old as her youngest brother Which expression could be used to show her age? A 2b - 1 B. 1 - 2b C 2b + 1
special ed, kid 1	X		
special ed, kid 2		X	

Viewing #1 of 15

Figure 5. Interface Design for the Teacher Perception Survey.

To judge who was eligible for each category—1%, 2%, or standard items—a simple majority rule was applied. In other words, the test including the majority number of items which the teacher rated on for the student was selected for that student. For example, if a teacher selected 3 1% items, 7 2% items, and 5 standard items for a student, the student is classified as 2% student from the Teacher Perception Survey.

Student performance test. In addition to the Teacher Perception Survey, all participating students with disabilities took the same 20 items comprised of the 10 2% items and 10 standard items that were presented to teachers for the Teacher Perception Survey in the Stage 1. Considering the endurance of students with disabilities, the test was broken down into test 1 and test 2 with 10 each items. Four out of 39 participating students with disabilities did not take both tests, so they completed 10 items rather than 20 items. These students took 5 2% items and 5 standard items.

Using the student performance data, students were scored as 'pass' or 'fail' for both standard items and 2% items based on the following rules: 1) students were scored as 'pass' if students responded correctly 5 or more out of 10 items; 2) students were scored as 'pass' if students completed only test 1 and responded correctly 3 or more out of 5 items; 3) otherwise, students were scored as 'fail'. Then using Table 3, participating students with disabilities were classified as 2% students or other level of students. For example, a student was classified as a 2% student if the student correctly answered 5 items out of 10 standard items, 6 items out of 10 2% items, and the student's teacher judged this student as 1% or 2% student. In the case that a student completed only test 1 (i.e., the student completed only 5 items of 2% items and 5 items of standard items), the student was identified as a 2% student when the student responded 3 items correctly in both standard items and 2% items and the result of Teacher Perception Survey was 1% or 2%. Based on the identification procedure, 23 students were identified as 2% students.

Table 3

Rules for Classifying Students as 2% Based on Student Performance Test and Teacher Perception Survey Results

Standard Items	2% Items	Teacher Perception Survey	ID Number	Number of Students
P	P	1% or 2%	11, 22*, 28, 38	4
P	F	1% or 2%	25, 30*, 33*	3
F	P	1% or 2%	3, 10, 18, 21, 26, 27, 29, 31, 34, 35, 36, 37, 40, 42	14
F	F	General Ed	7, 16	2
			Total	23

Note. * indicates id number of students who took only 5 items. General Ed= general education, P=pass, and F=fail.

Second, the same 20 items used in the identification of 2% students were delivered to general education students without being broken down into test1 and test2.

Data Analyses

First, the Item Response Theory (IRT) model was used to analyze the data. Among diverse IRT models, this study used the one parameter logistic (1PL) model, also known as the Rasch measurement model.

$$P\left(X_{is} = 1 \middle| \theta_s, \beta_i\right) = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}, \text{ where } i=1, 2, 3, \ldots, \text{n.} \qquad \text{Equation 1}$$

Equation 1 explains the probability that person *s* responds correctly to item *i*. The probability *P* is governed by the student ability, θ_s and the item difficulty, β_i. That is, given the student's ability, as an item becomes more difficult, the probability of a correct response decreases. Also, given an item's difficulty, as the student's ability increase, the probability of a correct response increases. The Rasch model is the most simplistic model in IRT since each item has only one parameter – the item difficulty β_i. In the Rasch model, all items have equal item discrimination and equal probability of correct guessing.

It was important to test statistically the hypothesis that item difficulty estimates are invariant across different grade levels (grades 6, 7, and 8) and different populations (2% students and general education students) because this study analyzed data from several groups and levels of population. To conduct this test, differential item functioning (DIF) analyses were used. Both MH and LR procedures were applied to detect uniform DIF, while the LR procedure was used to identify non-uniform DIF. Aligned with the research questions stated below, the following research hypotheses (for DIF analyses) were tested. Considering the small sample size of this study, the 1.5 of Mantel-Haenszel slice width (MHSLICE) instead of the default value, 0.1 was applied for the MH procedure. To investigate DIF, the sample is divided into different classification groups such as reference groups and focal groups to investigate DIF. Linacre (2006) illustrates that the MHSLICE specifies the width of the slice in logits of the latent variable to be included in each cross-tabulation. Linacre (2006) recommends large MHSLICE when sample size is small but an approximate Mantel-Haenszel estimate is wanted. In addition to the DIF techniques, Analysis of variance (ANOVA) was applied to compare mean differences across level of disability and grade. All analyses were conducted with an initial 0.05 level of the statistical tests of significance. Specific research hypotheses and data analyses per research questions are as follows.

Research Question 1.Are there grade levels at which the *item parameters* estimated from general education students in grades 6-8 are not different from those obtained using the 2% student sample?

Disability Level Item Parameter Invariance

1-(a) Are estimated *item parameters* of the alternate assessment invariant across level of disability when comparing 2 % students in grade 8 with general education students in grade 8?

Research Hypothesis 1-(a): There is no DIF in an item difficulty and test information between 2% students and general education students.

Data Analysis: To detect uniform DIF between 2% and general education students in grade 8, both the MH and the LR procedures were investigated. Also for identification of non-uniform DIF, the LR procedure was applied. The Bonferroni adjustment

($\alpha_{PC} = .005$) was applied to the results of both the MH and LR procedures to adjust the alpha for 10 items which were analyzed for this study.

Grade level Item Parameter Invariance

1-(b) Are estimated *item parameters* of the alternate assessment invariant across grade level when comparing general education students across grades 6, 7, and 8?

Research Hypothesis 1-(b): There is no DIF in an item difficulty and test information across the grades.

Data Analysis: To detect uniform DIF across the grades 6, 7, and 8 in general education students, both the MH and the LR procedures were investigated. Also to identify non-uniform DIF, the LR procedure was applied. The Bonferroni adjustment ($\alpha_{PC} = .00167$) was applied to the results of the MH procedure to adjust the alpha for 10 items and 3 group comparison for each item. In addition, $\alpha_{PC} = .005$ was applied to the results of the LR procedure for the Bonfferoni adjustment.

Interaction of Disability and Grade

1-(c) Do estimated *item parameters* depend on an interaction of disability and grade level when comparing 2% students in grade 8 and general education students in grade 6?

Research Hypothesis 1-(c): There is no interaction of disability and grade level in an item difficulty and test information.

Data Analysis: Both the MH and the LR procedures were investigated to identify uniform DIF. Also for the identification of non-uniform DIF, the LR procedure was applied. To adjust the alpha for 10 items, the Bonferroni adjustment ($\alpha_{PC} = .005$) was applied to results of both the MH and the LR procedure.

1-(d) Do estimated item parameters depend on an interaction of disability and grade level when comparing 2% students in grade 8 and general education student in grade 7?

Research Hypothesis 1-(d): There is no interaction of disability and grade level in an item difficulty and test information.

Data Analysis: To detect uniform DIF, both the MH and the LR procedures were investigated. Also the LR procedure was applied to identify non-uniform DIF. The Bonferroni adjustment ($\alpha_{PC} = .005$) was applied to the results of both the MH and the LR procedure to adjust the alpha for 10 items which were analyzed for this study.

Group Mean Differences

Research Question 2. Are there grade levels at which the estimated *ability* of general education students in grades 6-8 is not different from that of the 2% student sample in grade 8?

Research Hypothesis 2: Mathematics ability in general education students in grades 6, 7, and 8 will be higher than in 2% students in grade 8.

Data Analysis: One-way analyses of variance (ANOVA) were conducted to evaluate mean difference in mathematics ability across grades and between 2% students and general education students.

CHAPTER IV

RESULTS

This chapter presents the results of supplementing 2% student data sets with data obtained from general education students when developing mathematics alternate assessment items. Data analyses and results are organized by the primary research questions of this study. First, the descriptive statistics of participants are provided. Next, DIF analyses were applied to investigate disability level and grade level item parameter invariance. In addition, the interaction of disability and grade on responding to 2% items was also examined by DIF analyses. Finally, group mean differences across grades and between 2% and general education population were examined using ANOVAs.

Descriptive Statistics of Participants

Performance by 235 general education students in grades 6, 7, and 8 and 23 2% students in grade 8 were included in data analyses. The number of participants per group varied from 23 to 113. Out of 20 items participating students took, responses on 10 2% items (i.e., items 1, 3, 5, 7, 9, 11, 13, 15, 17, and 19) were analyzed for this study. The mean per group ranged from 5.87 to 7.61. Means and standard deviations for participants' performance on 10 2% items are presented in Table 4.

Table 4

Performance on 2% Mathematics Items by 2% and General Education Students

	2%			General Education		
Grade	M	SD	N	M	SD	N
6	--	--	--	6.49	2.00	113
7	--	--	--	6.79	1.98	52
8	5.87	2.01	23	7.61	2.21	70

Research Question 1

Are there grade levels at which the item parameters estimated from general education students in grades 6-8 are not different from those obtained using the 2% student sample?

(a) Are estimated item parameters of the alternate assessment invariant across level of disability when comparing 2 % students in grade 8 with general education students in grade 8?

(b) Are estimated item parameters of the alternate assessment invariant across grade level when comparing general education students across grades 6, 7, and 8?

(c) Do estimated item parameters depend on an interaction of disability and grade level when comparing 2% students in grade 8 and general education students in grade 6?

(d) Do estimated item parameters depend on an interaction of disability and grade level when comparing 2% students in grade 8 and general education student in grade 7?

To test statically the hypothesis that item difficulty estimates are invariant across different grade levels (grades 6, 7, and 8) and different populations (2% students and general education students), DIF analyses were conducted. Both uniform DIF and non-uniform DIF were investigated for this study. Not only the MH and but also the LR procedures were applied for detection uniform DIF, whereas the LR procedure was used for identification of non-uniform DIF.

MH Procedure for Identification of Uniform DIF

The MH procedure was used to detect uniform DIF. Table 5 presents the number of DIF items, percentage of DIF items, and DIF classification suggested by Zwick and Ercikan (1989). The DIF was regarded as moderate or large according to the absolute value of MH-Delta and the significance of the differential functioning statistical test. Absolute values over 1.5 and a statistically significant test at the .05 level indicate large DIF, while absolute values between 1 and 1.5 and a statistically significant test at the level .05 would be classified as moderate. The results identified 2 cases as displaying a large DIF and 3 cases with a moderate DIF. Specifically, item 13 displayed large uniform DIF ($|\Delta_{MH}|$=1.76) between general education students in grade 6 and students in grade 8. This item also presented large uniform DIF ($|\Delta_{MH}|$=1.95) between 2% students in grade 8 and general education students in grade 8. Item 7 was detected as displaying uniform DIF with moderate effect size ($|\Delta_{MH}|$=1.07) when comparing general education students in grade 6 with students in grade 7. Also, the comparison between general education students in grade 7 with students in grade 8 identified item 1 with moderate uniform DIF ($|\Delta_{MH}|$=1.09). Finally, item 11 presented moderate DIF ($|\Delta_{MH}|$=.1.17) between 2%

students in grade 8 and general education students in grade 6. After applying Bonferroni adjustment, only item 13 between general education students in grade 6 and general education students in grade 8 presented DIF.

Table 5

Levels of Grade or Disability by MH Procedure-Classification of Uniform DIF

Level		Number of DIF Items	% of DIF Items	DIF Classification	
				Moderate	Large
Grade	6-7 General Ed	1	10	1	-
	7-8 General Ed	1	10	1	-
	6-8 General Ed	1	10	-	1
Disability	2% grade 8 -General Ed grade 6	1	10	1	-
	2% grade 8 - General Ed grade 7	-	-	-	-
	2% grade 8 - General Ed grade 8	1	10	-	1

Note. MH = Mantel-Haenszel, DIF = differential item functioning, and General Ed = general education. After applying Bonferroni adjustment, only item 13 between general education students in grade 6 and general education students in grade 8 presented DIF. mhslice=1.5

LR Procedure for Identification of Uniform DIF

In addition to the MH procedure, the LR procedure was applied to detect uniform DIF. Table 6 presents the number of DIF items, percentage of DIF items, and DIF classification based on Jodoin's and Gierl's (2001) guideline. DIF was regarded as moderate or large based on the value of $R^2\Delta$-U, and the significance of the differential functioning statistical test. Values over .07 and a statistically significant test at the .05 level indicate large non-uniform DIF, whereas values between .035 and .07 and a statistically significant test at the level .05 would be classified as moderate. Three items (items 1, 3, and 13) presented uniform DIF between general education students in grade 6 and students in grade 8 (see Figure 6 for item 13). While two (items 1 and 13) of them showed moderate effect size ($R^2\Delta$-U=.043 and $R^2\Delta$-U=.055 respectively), one item (item 3) displayed large effect size ($R^2\Delta$-U=.110). In addition, uniform DIF occurred in one item (item 1) with the large effect size ($R^2\Delta$-U=.097) between 2% students in grade 8 and general education students in grade 8. As with the results of the MH procedure, after applying the Bonfferoni adjustment, only item 13 between general education students in grade 6 and general education students in grade 8 displayed uniform DIF.

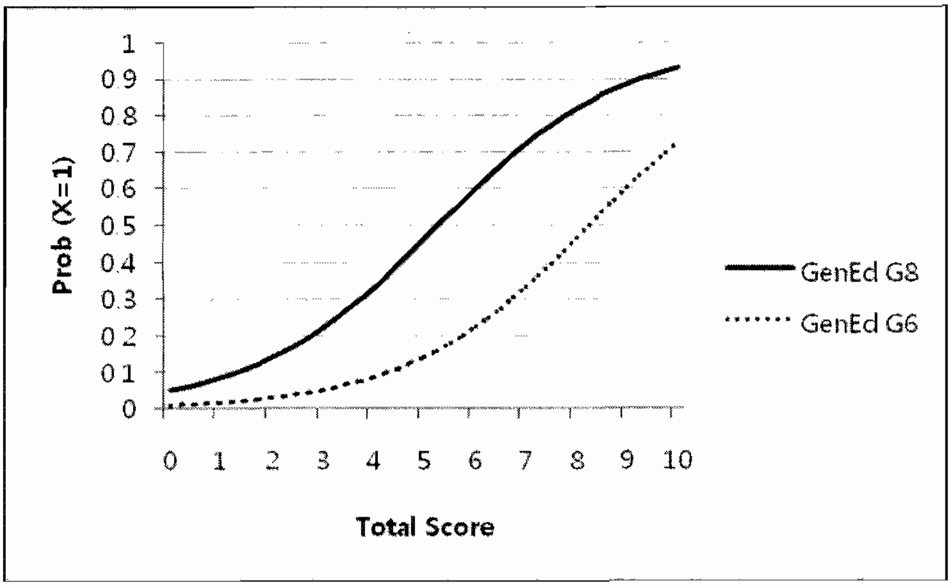


Figure 6. Item 13 with Uniform DIF—General Education Grade 8 and Grade 6.

Table 6.

Levels of Grade or Disability by LR Procedure-Classification of Uniform DIF

Level		Number of DIF Items	% of DIF Items	DIF Classification	
				Moderate	Large
Grade	6-7 General Ed	-	-	-	-
	7-8 General Ed	-	-	-	-
	6-8 General Ed	3	30	2	1
Disability	2% grade 8 -General Ed grade 6	-	-	-	-
	2% grade 8 - General Ed grade 7	-	-	-	-

2% grade 8 - General Ed grade 8	1	10	-	1

Note. LR = logistic regression, DIF = differential item functioning, and General Ed = general education. After applying Bonferroni adjustment, only item 13 between general education students in grade 6 and general education students in grade 8 presented DIF.

Consistency between the MH and the LR Procedures

Table 7 presents the number and percentage of items identified as showing uniform DIF by each procedure. In addition, Table 7 shows the number and percentage of uniform DIF items that were identified consistently by both MH and LR procedures. When comparing general education students in grade 6 with students in grade 8, 1 item was identified as displaying with uniform DIF by the MH procedure, whereas 3 items were detected as showing uniform DIF by the LR procedure. One item (item 13) was consistently identified as showing DIF by both MH and LR procedures even though the same item was classified as displaying different effect size of DIF (i.e., large DIF by MH procedure, and moderate DIF by LR procedure). The associated matching percentage was 33%. The comparison 2% grade 8 with general education grade 8 by both MH and LR procedures presented one DIF item. However, the DIF item by MH procedure was item 13, whereas it was item 1 by the LR procedure.

Table 7

Consistency between Procedures for Uniform DIF

Level		Number of DIF Items (%)		Consistency
		MH	LR	
Grade	6-7 General Ed	1(10%)	-	-
	7-8 General Ed	1(10%)	-	-
	6-8 General Ed	1(10%)	3(30%)	1(33%)
Disability	2% grade 8 -General Ed grade 6	1(10%)	-	-
	2% grade 8 - General Ed grade 7	-	-	-
	2% grade 8 - General Ed grade 8	1(10%)	1(10%)	0(0%)

Note. MH = Mantel-Haenszel, LR = logistic regression, DIF = differential item functioning, and General Ed = general education.

LR Procedure for Identification of Non-uniform DIF

In order to detect non-uniform DIF, LR procedure was applied. Table 8 presents the number of non-uniform DIF items, percentage of non-uniform DIF items, and DIF classification based on Jodoin and Gierl (2001). Item 17 was detected as displaying large non-uniform DIF in both the comparison between general education grades 6 and 8 and the comparison between 2% grade 8 and general education grade 8. The DIF effect sizes were $R^2\Delta$-U=.307 and $R^2\Delta$-U=.244 respectively. In addition, item 13 showed non-uniform DIF with large effect size ($R^2\Delta$-U=.098) when comparing 2% grade 8 with general education grade 8 (see Figure 7). After applying the Bonfferoni adjustment, no DIF item was detected from all cases.

Table 8.

Levels of Grade or Disability by LR Procedure-Classification of Non-uniform DIF

Level		Number of DIF Items	% of DIF Items	DIF Classification	
				Moderate	Large
Grade	6-7 General Ed	-	-	-	-
	7-8 General Ed	-	-	-	-
	6-8 General Ed	1	10	-	1
Disability	2% grade 8 -General Ed grade 6	-	-	-	-
	2% grade 8 - General Ed grade 7	-	-	-	-
	2% grade 8 - General Ed grade 8	2	20	-	2

Note. LR = logistic regression, DIF = differential item functioning, and General Ed = general education. After applying the Bonfferoni adjustment, no DIF item was detected.

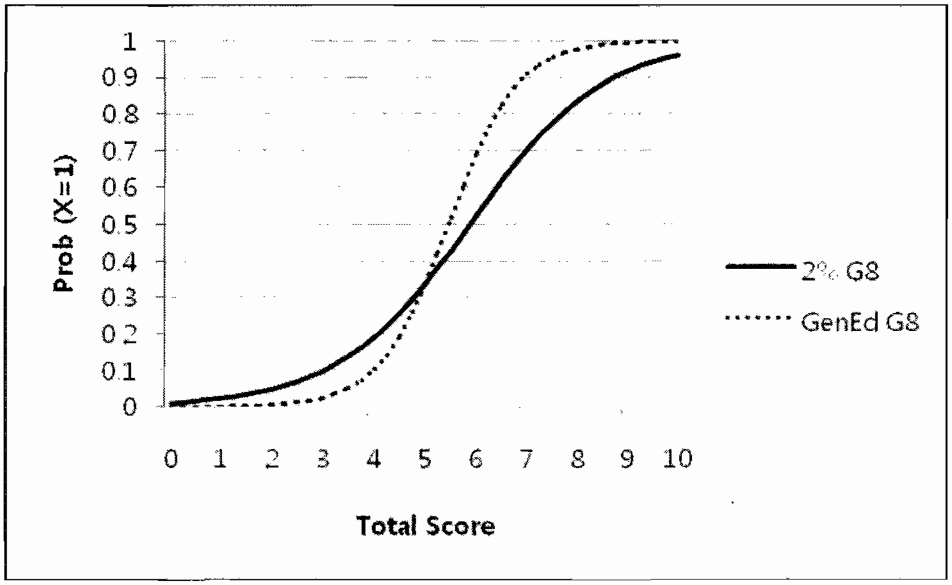


Figure 7. Item 13 with Non-uniform DIF—2% Grade 8 and General Education Grade 8.

Test Information

A test, as comprised of any set of items, has an aggregated test information function $TI(\theta)$ for examinee with ability θ (i.e., $TI(\theta) = \sum_{i-1}^{n} I_i(\theta)$). A test information function can be established with estimated item parameters. The test information curves indicate two test features— (1) location where the test is most informative, and (2) amount of information at various scale points. The test information function is essential in determining how well a test is performing because the test information function has a direct relationship with standard error (SE) of ability estimation $SE(\hat{\theta})$, i.e., $SE(\hat{\theta}) = \frac{1}{\sqrt{TI(\theta)}}$ (Embretson & Reise, 2000). Figure 8 illustrates test information and SE curves for 2% students in grade 8 and general education students in the same grade. This comparison did not show any item-level DIF items after applying the Bonfferoni adjustment. Test information curves for both groups showed maximum information at the same scale location. Also the figure presented that the amount of information was very similar at the point of maximum information even though the test information for 2% students seemed relatively higher than that for general education students in the same grade. So the test information curves appeared consistent with the DIF findings indicating invariance with respect to the item difficulty parameters.

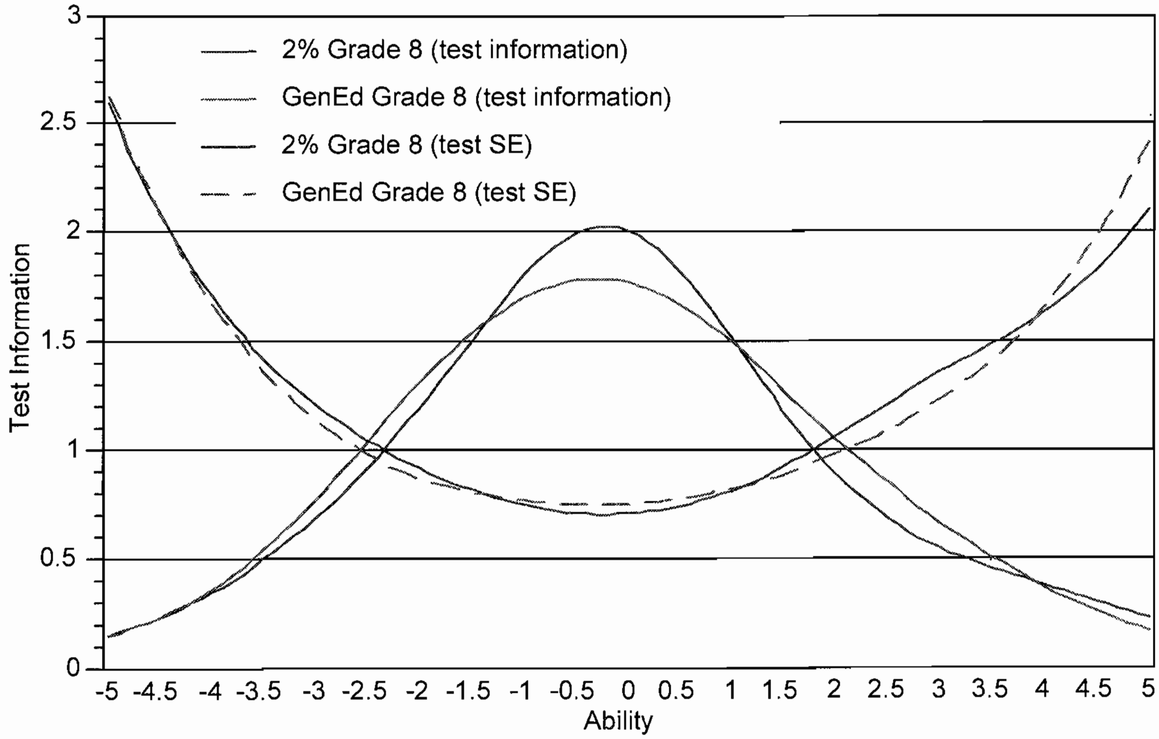


Figure 8. Test information and SE curves for 2% students and general education students in grade 8.

Comparison between general education students in grades 6 and 8 displayed a large uniform DIF item even after applying the Bonfferroni adjustment. Figure 9 illustrates the test information and SE curves for these groups. The two test information curves appeared identical, indicating that the tests are comparably informative and so there is no test-level DIF even though there was an item-level DIF.

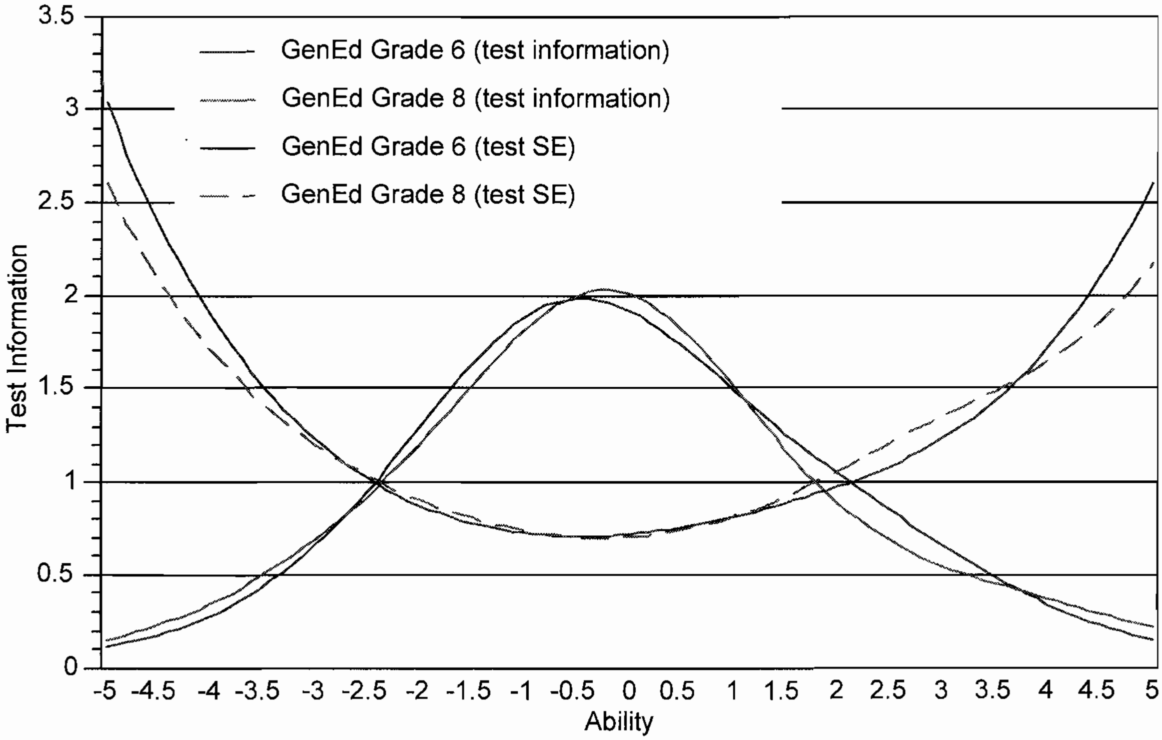


Figure 9. Test information and SE curves for general education students in grades 6 and 8.

Research Question 2

Are there grade levels at which the estimated ability of general education students in grades 6-8 is not different from that of the 2% student sample in grade 8?

One-way analyses of variance were conducted to evaluate mean differences in mathematics ability across grades and between 2% students and general education students.

General Education Students in Grades 6, 7, and 8

Table 9 shows that there was a significant effect of levels of grade on performance of 2% items, $F(2, 232) = 6.54, p < .05$. To determine which grade levels among grade 6, 7, and 8 showed significant mean differences, multiple comparison using Sidak procedure at .05 significance level was applied. The post hoc analysis determined that there were significant differences between grades 6 and 7 as well as between grades 6 and 8. However, there was no significant difference in the performance on 2% items between grade 7 and grade 8.

Table 9

Analysis of Variance for General Education Students in Grade 6, 7 and 8 on 2% Items

Source	*df*	*SS*	*MS*	*F*
Grade	2	55.64	28.82	6.54*
Error	232	987.49	4.26	
Total	234	1043.12		

*p < .05.

2% and General Education Students in Grade 8

Table 10 presents results indicating statistically significant difference in performance on 2% items were observed between 2% students and general education students in grade 8, $F(1, 91) = 11.12, p < .05$. General education students in grade 8 performed significantly higher than 2% students in grade 8 on 2% items.

Table 10

Analysis of Variance for 2% and General Education Students in Grade 8

Source	*df*	*SS*	*MS*	*F*
Disability	1	52.70	52.70	11.12*
Error	91	431.19	42.74	
Total	92	483.89		

*$p < .05$.

2% 8th Grade and General Education 7th Grade

There was no significant mean difference between 2% students in grade 8 and general education students in grade 7, $F(1, 73) = 3.33, p > .05$. That is, there was no significant effect of level of disability or grade on performance of 2% items,

Table 11

Analysis of Variance for 2% 8th Grade and General Education 7th Grade Students

Source	*df*	*SS*	*MS*	*F*
Grade/Disability	1	13.46	13.56	3.33
Error	73	295.28	4.05	
Total	74	308.75		

*$p < .05$.

2% 8th Grade and General Education 6th Grade

There was no significant mean difference between 2% students in grade 8 and general education students in grade 6, $F(1, 134) = 1.79, p > .05$. The result indicated that there was no significant effect of levels of disability or grade on performance of 2% items

Table 12

Analysis of Variance for 2% 8th Grade and General Education 6th Grade Students

Source	*df*	*SS*	*MS*	*F*
Grade/Disability	1	7.28	7.28	1.79
Error	134	544.74	4.07	
Total	135	532.12		

*$p < .05$.

Summary

Ten items created for 2% students in grade 8 were analyzed for this study. Results of the DIF analyses indicated that overall, estimated item parameters were invariant across grades and disability levels. First, all estimated item parameters, except one item with large uniform DIF (item 1 by the LR procedure and item 13 by the MH procedure), were invariant across level of disability when comparing 2% students in grade 8 with general education students in grade 8. The LR procedure also detected 2 items as displaying large non-uniform DIF: items 13 and 17 (see Appendix D for the specific item). Second, the MH procedure indicated that item parameters of 2% items were invariant across grades when comparing general education students in grades 6 and 7 except one item with moderate uniform DIF (item 11). Also, only one item (item 1) with moderate uniform DIF was identified by the MH procedure when comparing general education students in grades 7 and 8. Comparison between grade 6 and grade 8 detected items as displaying either uniform DIF or non-uniform DIF. Specifically, one large uniform DIF was detected by the MH procedure (item 13), while 2 moderate uniform DIF items (items 1 and 13) and one large uniform DIF item (item 3) were identified by the LR procedure. In addition, the LR procedure detected one item with large non-uniform DIF (item 17). Third, the MH procedure identified one moderate uniform DIF (item 11) when comparing 2% students in grade 8 with general education students in grade 6. Fourth, the DIF analyses by the both MH and LR procedures indicated that estimated item parameters were not dependent on an interaction of disability and grade level when comparing 2% students in grade 8 with general education students in grade 7.

Finally, after applying the Bonfferoni adjustment to all above cases, only item 13 presented uniform DIF from both the MH and the LR procedures when comparing general education students in grade 6 with general education students in grade 8. Not only item-level DIF but also test-level DIF was examined. The test information curves indicated that there was probably not test-level DIF. Even when there was DIF item after applying the Bonfferoni adjustment, it had little apparent impact on the test information.

One-way ANOVAs were conducted to document group mean differences across disability and grade levels. The results of one-way ANOVAs first showed that there were significant group mean differences across disability level. That is, general education students in grade 8 performed significantly higher than 2% students in the same grade. In addition, general education students in grade 6 performed significantly lower than students in grade 7 and students in grade 8. Furthermore, the results of ANOVA indicated no significant differences in mathematics performance on 2% items between general education students in grade 7 and students in grade 8. Finally, 2% students in grade 8 showed comparable mathematics ability on 2% items to general education students in both grade 6 and grade 7.A formal discussion of these results and their implications for practice and future research is provided in the chapter that follows.

CHAPTER V

DICUSSION

No Child Left Behind (NCLB; 2001) requires states to test all students from grades 3 through 8, including students with disabilities. Students with disabilities often have a difficulty performing on tests specifically designed for general education students. In order to accurately and validly determine skills of students with disabilities, modifications to the design of state assessments may be warranted to address their unique learning difficulties. Designing alternate assessments requires psychometric modeling for assurances that the measurement tool is technically adequate and appropriate for the intended purposes (U.S. Department of Education, 2007). Part of this design modification would be determining the adequacy of item-level functioning, which is a challenge considering the small numbers of students with these learning difficulties.

The present study focused on addressing this sampling issue for students with disabilities, referred to as "2%" students. Two percent students are students with disabilities for whom the regular state assessment is considered too difficult, yet the alternate achievement standards are too easy. This study investigated whether psychometric characteristics of mathematics alternate assessment items created for 2% students in grade 8 can be meaningfully estimated with data obtained from general education students in grades 6, 7, and 8. DIF analyses and ANOVAs of students' performance revealed the possible benefit of supplementing 2% student data sets with data obtained from general education students in lower grades as a possible method of addressing the sampling issue in test development of alternate assessment items. The remainder of this chapter summarizes the major findings by focusing on two areas: (1) possibility and utility of using lower grade general education students for estimating alternate assessment item parameters; and (2) potential reasons for DIF occurrences. Next, the study's limitations are provided. Finally, recommendations for future research on designing alternate assessments are provided.

Possibility and Utility of Using General Education Student Data

There is agreement among educational researchers that large sample sizes are beneficial in test development and validation. Unfortunately, obtaining adequate sample sizes is not always possible (Rabinowitz & Sato, 2006; Muniz, Hambleton, & Xing, 2001). Low prevalence of the targeted population definitely causes a sampling issue, when developing alternate assessments for 2% students,

The findings of the present study documented the possibility and utility of using lower grade general education student samples in item validation. Overall, item response patterns of 2% students in grade 8 on 2% items were similar to those of general education students in grades 6 and 7. Specifically, only one item with moderate uniform DIF was detected by the MH procedure when comparing 2% students in grade 8 with general education students in grade 6. Even this item did not show DIF after applying the Bonfferoni adjustment for Type I rates—incorrectly rejecting the null hypothesis (i.e., P(X=1|θ) is constant for both groups). In addition to the MH procedure, the LR procedure was used to investigate non-uniform DIF as well as uniform DIF. The LR procedure did not identify any DIF items. The DIF analyses with both the MH and LR procedures indicated that the estimated item characteristics obtained with 2% students in grade 8 were comparable to those obtained with general education students in grade 6. Additionally, item parameter estimations using 2% students in grade 8 and general education students in grade 7 did not display any DIF items using both the MH and LR procedures. This indicated that the item response patterns of general education students in grade 7 were also similar to those of 2% students in grade 8.

The ANOVA results also supported using general education student data for developing alternate assessment for 2% students. The mean of 2% students' performance (M=5.87), though lower than general education 6th grade students (M=6.49), was not significantly different and in fact comparable. Similar results were found when comparing 2% students to 7th grade students (M=6.79). Not surprisingly, when comparing 2% students to their non-disabled grade-level peers (M=5.87 and M=7.61, respectively), a significant difference was detected. This result is consistent with the generally accepted belief that 2% students may be functioning 2 grade levels below their peers without disabilities (Thurlow et al., 2002).

Reasons for DIF

This study intended to examine which grade level of general education students is comparable with 2% students in grade 8 in terms of item response pattern. While generally not detected, DIF was present in some comparisons. Considering the complication of level of disability and grade used in this study, the hypothesis was that there was no DIF across grades and disability level. The unidimensional measurement construct of interest is ability (Embretson & Reise, 2000). The presence of DIF (i.e., the items do not function the same between or among groups) suggests that construct-irrelevant item response variance related to student characteristics may be present resulting in estimation of item characteristic that are not sample invariant (Embretson & Reise, 2000).

Cognitive Complexity of the Item

A possible reason for DIF occurrence is related to item characteristics, i.e., cognitive complexity of the item. Cognitive complexity was initially defined as "level of thinking" in the psychology field (McDaniel & Lawrence, 1990). Research on cognitive complexity has focused on how cognitive complexities affect measuring students' performance (Webb, 1999; 2002; 2007; Bloom, 1956; Krathwohl, 2002; Gorin & Emretson, 2006; Gorin, 2006). This study focused on 2% students who are in regular classrooms and are learning grade-level content but are not expected to reach grade-level achievement standards within an academic year because of their disability. However, according the NCLB 2007 regulations, 2% students must be assessed on the same grade-level content as their non-disabled peers. One potential method of determining a student's content skills while avoiding bias due to their disability is to reduce the cognitive complexity of the items. Prior to this study, mathematics items were developed to reduce the cognitive complexity, and this project analyzed the validity of those items.

The findings of the present study indicated that cognitive complexities of the item would introduce construct-irrelevant variance affecting item response behaviors and overall test score. Specifically, 2% students in grade 8 differently responded to item 13 from general education students in the same grade. Also, non-uniform DIF items (items 13 and 17) seemed to imply that there were problems in these items because these items did not fully reflect these students' skills. Taking a close look at the DIF items, all DIF items detected between 2% students and general education students were measuring skills in algebra. The two items (items 13 and 17) were developed to align with different content standards (see Appendix A), and had students interpreting a graph demonstrating a linear equation. Krathwohl (2002), in his discussion on cognitive complexity, would categorize these items as both measuring procedural knowledge. These findings may imply that the modifications made to the graphing items failed to reduce the cognitive complexity of the item and may not adequately determine the 2% students' skill in algebra. Further discussion on this topic is provided in the recommendations for future research section later.

Content Exposure

Another possible reason for the identified DIF occurrences could be related to what the 2% students have been taught or content exposure. Content exposure may result in DIF, especially when considering the variability in content standards across grade levels. As previously mentioned, the 2% items must be designed to assess 8th grade content standards, regardless of what content the students have been taught. Even though general education students in grades 6 and 7 are expected to have better mathematics skills than 2% students in grade 8, some of content standards in grade 8 may not be introduced to students in lower grades. When comparing 2% students in grade 8 and general education students in grade 6, one item (item 11) was identified as displaying moderate uniform DIF. Although the DIF disappeared after applying the Bonfferoni adjustment, the result seemed to imply that general education students in grade 6 may respond differently from 2% students in grade 8 potentially because the 6th grade students were not taught the same content (i.e., linear algebra). This content exposure may also be relevant to interpreting the analyses between student performance in grades 6 and 8 where the LR and MH analyses detected DIF. Even after applying the Bonfferoni adjustments to item 13, large to moderate DIF were detected on both the MH and LR procedures, respectively. In future studies, the issue of content exposure should be more fully addressed to more accurately understand item development and validation.

Limitations

The findings of this study need to be considered within the context of its limitations which included a limited sample of 2% students, issues in student identification/teacher recruitment, the Teacher Perception Survey, and test length. Each issue is discussed in this section.

Sample Size

The first and the biggest limitation of the current study is the small sample size of the 2% student (N=23), especially considering this study applied DIF analyses. Swaminathan and Rogers (1990) argue that a small sample may jeopardize the stability of the estimate of odds ratio in each group in the MH procedure. Additionally, a small sample limits the power of the LR procedure through its effect on estimation of item parameter, resulting in doubts about using test statistics (p value) as a valid indicator to detect DIF. It is desirable, even for the MH procedure (specifically designed to address small samples), to have at least 200 students per group to assure adequate detection rates and prevent Type I errors (Swaminathan & Rogers, 1990). Researchers advise that sample sizes of less than 100 may weaken results with the MH procedure even when the substantial DIF is detected (Mazor, Clauser, & Hambleton, 1992; Parshall & Miller, 1995). Even though Muniz, Hambleton, and Xing (2001) address the possibility of using small sample size with the MH procedure, the smallest sample size used in their study was 100—50 persons in the reference group and 50 persons in the focal group. As noted, this study included only 23 students in the 2% group (i.e., focal group). In this study, the Bonfferoni adjustment to reduce the Type I errors was applied, but considering the sample size these results must be viewed with caution and beg for replication in future studies (Muniz, Hambleton, & Xing, 2001).

Teacher Recruitment and Student Identification

A second limitation is related to the method and procedure for identifying which students were considered “2% students.” When recruiting 2% students for participation in the study, teachers were asked to identify 8th grade students who they believed best fit the profile of a “2% student.” Even though teachers were paid for their efforts, initial payment levels did not produce many teachers and students for participation. To increase numbers, the incentive was doubled to teachers and resulted in identifying more students and teachers. The vague criteria for what constitutes a student meeting "2% profile" limits how the obtained results can generalize to other states which use a different definition/criteria/method for identifying 2% students. Additionally, teachers may have overly identified students as 2% due to the monetary incentive provided (i.e., 16 students were later excluded because they did not meet actual 2% criteria). Student identification issues were also found on the Teacher Perception Survey.

Teacher Perception Survey

Identification of 2% students was based on the Teacher Perception Survey and the Student Performance Test. The Teacher Perception Survey consisted of 15 sets of 1%, 2%, and standard items. Unfortunately three out of 15 sets in the Teacher Perception Survey included only 1% and standard items. Having only 12 2% items in the Teacher Perception Survey may have affected the result of the identification of 2% students because the majority rule was applied for the identification of 2% (i.e., the test including the majority number of items which the teacher rated on for the specific student was selected for that student). The identification of 2% was not only dependent on the results of the Teacher Perception Survey but also incorporated the results from Student Performance Test. However, missing 3 2% items in the Teacher Perception Survey may affect the classification of 2% students and eventually challenge conclusions.

Test Length

A final limitation was that the test was comprised of only ten items. Swaminathan and Rogers (1990) address the importance of test length in DIF analyses: "the longer the test, the more reliable the total score" (p. 365). The authors argue that a more reliable total score (i.e., longer test length) would be required to get enhanced estimates of the parameters, because total score is used as a predictor in the LR procedure and as the criterion for grouping examinees in the MH procedure.

Recommendations for Future Research

Based on results and limitations of this study, several directions for future research are suggested. First, subsequent research should be conducted with larger sample sizes. Studies with larger sample sizes would provide (a) improved estimation of item parameter and (b) greater statistical power in data analyses.

Second, further investigation into the criteria of identification of 2% students is necessary to corroborate the findings from this study. 2% eligibility has been an issue since the final NCLB regulations announced because they only indicated that the eligibility will be determined by a student's IEP team. The eligibility criteria of 6 states reported by Lazarus et al. (2007) differed across each of the 6 states. The current study also established a different rule for 2% student identification. Defining and keeping stable eligibility criteria for the participating students in alternate assessments should be considered prior to designing an alternate assessment (Sato et al., 2007). In addition, there is a significant need to conduct research on development of specific guidelines or checklist for identifying the 2% students.

Third, item development for 2% students may benefit from careful consideration and systematic variation of cognitive complexity. The four categories of cognitive complexity are factual, conceptual, procedural, and metacognitive knowledge (Krathwohl, 2002). It is essential to complete research investigating 2% items directly addressing each level of cognitive complexity rather than only focusing on factual knowledge, which has been most common approach.

Fourth, in order to find out the different item response pattern between 2% students and general education students, not only the analysis on the correct answer but also the analysis on the distractors would be needed. The item response pattern on the distractors would provide a better understanding of 2% students' unique cognitive processing.

Fifth, future studies should estimate IRT models other than the 1PL Rasch model limited to item difficulty. Investigating 2PL or 3PL IRT models may provide more information about complicated item response patterns of 2% student, e.g., item discrimination and item guessing.

Finally, future studies need to focus more on why DIF occurs. As Gierl (2005) claims, there is lack of research addressing the interpretation of DIF, while substantial research has been conducted on how to detect DIF. Items showing DIF for two or more groups are considered to violate the unidimensionality assumption and large DIF would be a sign that the item is measuring an additional, unintended construct (i.e., construct-irrelevant variance). Future research should include not only the identification but also the interpretation of the construct that elicits group differences.

Conclusion

A significant challenge in developing valid assessments for students meeting the 2% criteria is demonstrating appropriate approaches and methodologies, considering their small numbers. This study investigated whether psychometric characteristics of mathematic alternate assessment items created for 2% students in grade 8 can be meaningfully estimated with data obtained from general education students in grades 6, 7, and 8. Participants included 23 2% students in grade 8 and 235 general education students in grades 6-8. Twenty three 2% students were identified through the Student Performance Test (10 standard items and 10 2% items) and the Teacher Perception Survey. Performance on 10 2% items by the 2% students and the general education students were analyzed using DIF techniques and ANOVAs to determine item-level functioning as well as test-level differences and group mean differences across grade level and student type (2% or general education students). The results from the present study indicated that the item response patterns of 2% students in grade 8 were comparable to those of general education students in grades 6 and 7. Additionally, 2% students in grade 8 showed comparable mathematics performance on 2% items when compared to general education students in grades 6 and 7. Considering the content exposure of students in lower grade levels, this study concluded that data from general education students in grade 7 would be more appropriate to be used in designing alternate assessment for 2% students in grade 8 than data from students in grade 6. The general conclusion is that using data obtained from general education students in lower grade levels may be an appropriate and efficient method of designing alternate assessment items.

APPENDIX A

FOCAL POINT STANDARD, OBJECTIVE, AND ITEM ALIGNMENT

Focal Point Standard	Objective	2% Item
Algebra	Use linear functions and equations to represent, analyze and solve a variety of problems and to make predictions and inferences.	Item1, Item 11, Item 17, Item 19
	Translate among verbal, tabular, graphical, and algebraic representations of linear functions.	Item 7, Item 13
	Recognize how the properties (i.e., slope, intercepts, continuity, and discreteness) of linear relationships are shown in the different representations.	
	Determine the slope of a line and understand that it is a constant rate of change.	
	Use systems of linear equations in two variables to represent, analyze, and solve a variety of problems.	
	Relate systems of two linear equations in two variables to pairs of lines that are intersecting, parallel, or the same line.	
Geometry and Measurement	Use properties of parallel lines, transversals and angles to solve problems including determining similarity or congruence of triangles.	
	Use similar triangles to find unknown lengths.	Item 3, Item 15
	Use models to show that the sum of the angles of any triangle is 180 degrees and apply this fact	

	to find unknown angles.	
	Use models to explore the validity of the Pythagorean Theorem using a variety of methods.	
	Analyze and apply the Pythagorean Theorem to find distances in a variety of 2- and 3-dimensional contexts.	
Data Analysis, Number and Operations, Algebra	Use descriptive statistics, including mean, median, mode, and range to summarize and compare data sets.	
	Organize and display data to pose and answer questions including pie charts, histograms, box plots, and scatter plots.	Item 5, Item 9
	Compare descriptive statistics and evaluate how changes in data affect those statistics.	
	Describe the strengths and limitations of a particular statistical measure and justify its use in a given situation.	
	Interpret and analyze graphical displays of data and descriptive statistics.	

APPENDIX B

ITEM-WRITING GUIDELINE FOR 2% MATHEMATICS ITEMS

BRT Item Writing 2% IES Math Items

Our goal is to have math items aligned with grade level content standards that are appropriate for use with students with cognitive disabilities. The students taking these math tests will be among the lowest 3% academically of students in their grade level. These students will be receiving special education services and are likely to need significant support to be able to function in inclusive (mainstreamed) classrooms.

When you write math items, keep in mind that we are trying to reduce the cognitive complexity of the tasks (while still retaining integrity in how they are tied to grade level content standards). Several of our leading researchers are currently working on a 'white paper' to describe what is meant by 'reducing cognitive complexity', but in the meantime, here are some basics to consider:

Pick the approach with the least manipulation needed

There are typically several possible ways in which one might structure / depict math operations. For example, in addition, you can represent the same problem in the following three ways (among others):

Item A	$\begin{array}{r} 31 \\ \underline{+\ 22} \end{array}$
Item B	31 + 22 = ___
Item C	Thirty-one plus twenty-two equals

Each approach differs in the amount of cognitive processing that is required to perform the operation, with the first option requiring the least amount of manipulation in the process.

For our tests, we want to use the option with the least amount of manipulation possible to get at the underlying skill. Thus, we will consistently use items written like Item A, above. The same basic concept applies across all types of calculations problems. We want to represent them in the format that requires the least amount of manipulation.

Address the standard, but do so simply

When selecting what numbers to use in a particular math problem, select numbers that are easier to work with as the student uses them to demonstrate mastery of the content standard.

Select words with care

For the math test, the words used in the questions / response choices should not get in the way of students ability to demonstrate their knowledge. Use simple language: short words; short declarative sentences.

These tests are real

The items you are writing will be used in several different states as part of actual large scale assessments for students with persistent learning difficulties. Please keep this in mind as you write: quality is more important than speed; the students deserve our best work.

IES General Item Writing Guidelines

Overall item writing goal - Focus students' attention on a single idea

Some general guidelines to help

- Understand the material before you start

- Think about the students you are writing for
- Be as clear and concise as possible
- Avoid irrelevant language, clues and difficulties

Specific things to keep in mind

- Address key verbs in the standards (recall, analyze, construct, recognize)
- Include all needed information in the question, so that answer choices are as simple as possible (X = poor example, O = good example)

X Bears O The article says that Bears are

a. Are omnivorous mammals a. Omnivores

b. Lay eggs like reptiles b. Herbivores

c. Do not like many vegetables c. Carnivores

- Keep grammar parallel between the question and each answer option

X Pedro left early because O Pedro left early because

a. He had to practice his dance. a. He had to practice his dance.

b. When he went home to eat.b. He had to go home to eat.

c. Why he met his new friend. c. It was time to meet his friend.

- Avoid 'all of the above' and 'none of the above' options
- Avoid negatives in questions and answers, especially double negatives
- Keep answer choices similar in length and complexity
- Make sure answer choices are mutually exclusive

X 45.80 is between O 45.80 is between

a. 45 and 46 a. 45 and 46

b. 40 and 50	b. 48 and 50
c. 40 and 80	b. 458 and 4580

Adapted from:

Michigan State University Scoring Office. Writing Test Items.

Retrieved 17 Aug 2007: http://www.msu.edu/dept/soweb/writitem.html

Frary, Robert B. (1995). More multiple-choice item writing do's and don'ts.

Retrieved 17 Aug 2007: http://PAREonline.net/getvn.asp?v=4&n=11

IES Important Item Writing Points (MATH)

- Use the excel files you receive to complete your work and do not change the name or extension (.xls) of the files. For each item set, you will be working on three excel files: one for each grade level in the set. Please think about the ways in which items should progress in difficultly as students move from one grade to the next, and apply this increase in difficulty as you write the items across the grade bands.
- Create **multiple-choice** test items to **address the standards** on your assignment.
- Examples are provided as starters, but you may use any kind of question you can imagine that addresses the standards.

- Study the standards, both the general (at the top of the spreadsheet) and the specific (in the Standard column) objectives. Make sure you understand what each one requires.

- Plan what your items might look like before you start to write. It may help to think of the types of items that will address each standard first, and then plan out how many of each you will write.

- Make items as **simple** and **direct** as possible, in the **most basic form** of the standard requirements.

- Always **reduce complexity** as much as possible while still addressing the standards.

- Maintain **appropriate language, vocabulary, background knowledge and interest** to students of the target grade level.

    - Do not use first person ("I") for story problems. Do not use topics of adult concern (insurance, marriage, etc.). Do not use the actual language of the standards unless necessary or appropriate.

- **Simplify language** as much as possible. Avoid idioms, long words, passive voice, and unnecessary clauses. More, shorter sentences are better than long, complex ones (see example page).

    - Use the EDL Core Vocab list to check the approximate grade level of a word. For all math items involving words, aim for vocabulary a minimum of 2 grade levels below the grade level of the items you are writing.

- Write **original** items. You may use any source for inspiration, ideas, or information but make sure your items are original.
    - Refer to the TX SDAA-II, textbooks, or even a simple Google search for ideas if you get stuck. http://www.tea.state.tx.us/student.assessment/resources/release/sdaa/index.html

- Make sure that each item stands alone and can be answered independently of all others.
- Adhere to the distribution of items among the standards, as on your template.
- Provide **three answer choices** for each item.
- Put the **correct answer** as the first answer choice (**A**) (they will be randomized later).
- Keep the three answer choices similar in length and complexity, differing only in content.
- Keep incorrect answer choices **relevant** to the problem. Do not put completely unrelated words or numbers as incorrect answer choices.
- We will be working with both a graphics designer and a web designer to move the raw items you create onto a computer platform for use with students in April. Because so many different people will be working with the items, it is essential that we all understand and **use the appropriate 'code' to let the graphics and web**

designer know what the final item should look like. See page 8 for a key to the code we will be using.

- Sketch, cut and paste, or describe your graphics on a separate paper. Put all graphics for an item on one sheet (see Graphics Sheet Example). Label them and mail or bring them to us. FotoSearch has lots of clip art to use as examples: http://www.fotosearch.com/

    - Check your email daily for updates, notes and further guidelines.

IES Reducing Language Complexity (MATH)

We need to keep minimal levels of language complexity so that students can better demonstrate their math skills, regardless of reading difficulties. There are several ways to keep the language complexity to a minimum. Here are a few examples:

1) Reduce unfamiliar words, long words, and idioms

X: Max is in charge of the raffle.

O: Julia is in charge of the bake sale.

X: Circle the clumps of eggs in the illustration.

O: Find the groups of eggs. Draw circles around the eggs.

2) Reduce subordinate and conditional clauses, passive voice, and logical connectors

X: Because the box was a cube with six equal sides, Jen calculated the area by…

O: The box is a cube with six equal sides. How did she get the area?

X: Two pencils in the box were found to be broken.

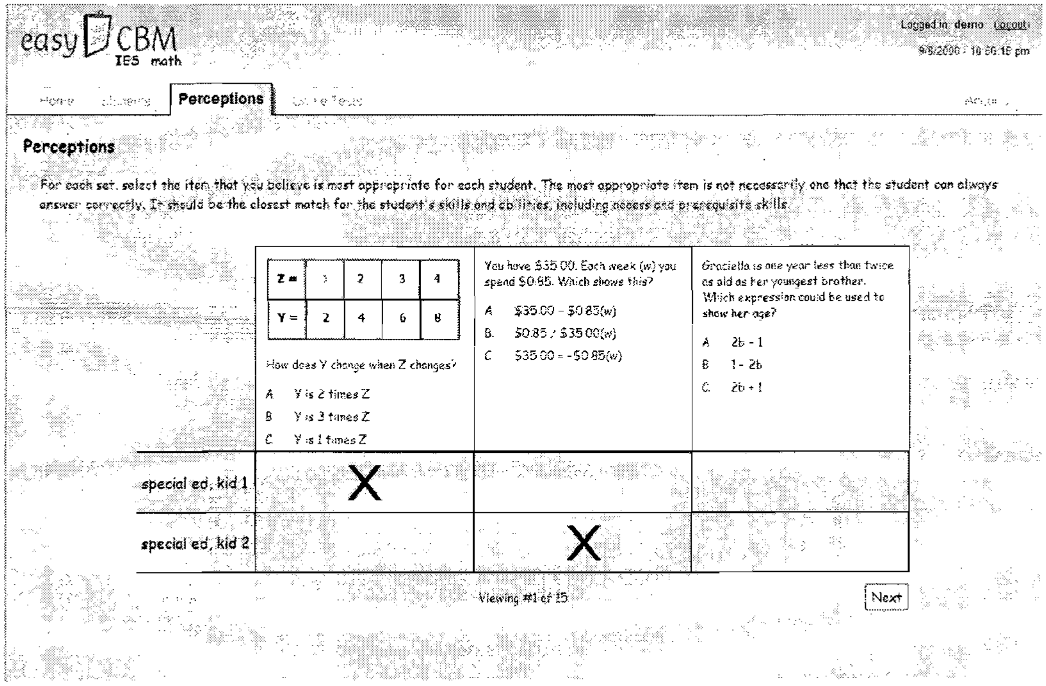
O: Dan found two broken pencils in the box.

X: If rulers cost $1.23 each, including tax, and Fred has $9.00, how many can he buy?

O: Rulers cost $1.23 each. Fred has $9.00. How many can he buy?

APPENDIX C

SAMPLE OF TEACHER PERCEPTION SURVEY

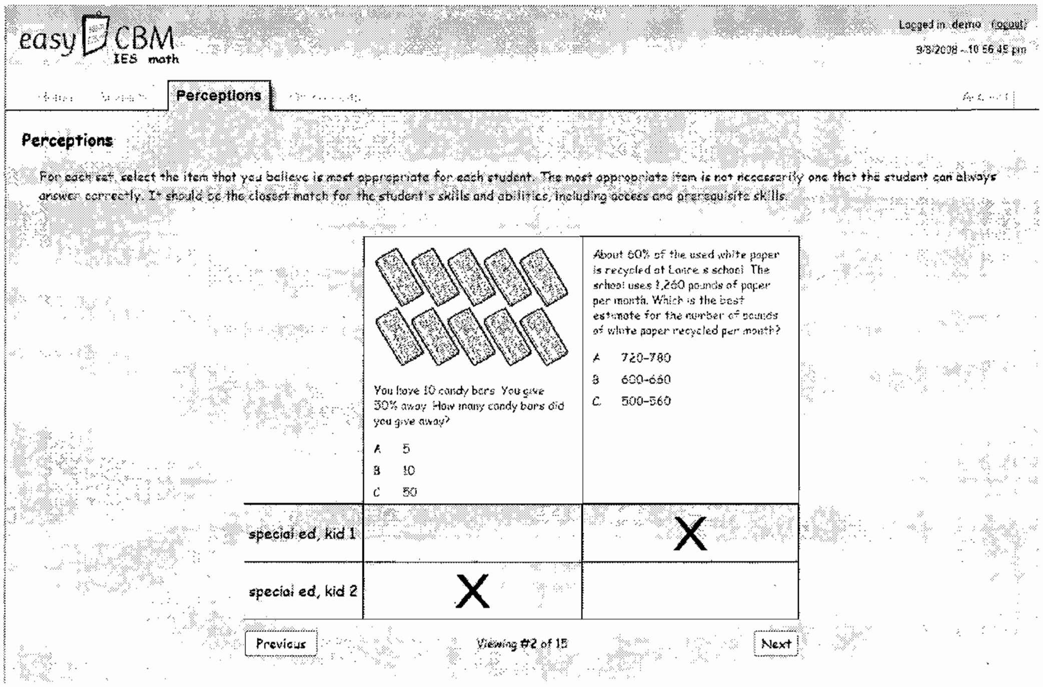


Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

	Z = 1 2 3 4 Y = 2 4 6 8 How does Y change when Z changes? A Y is 2 times Z B Y is 3 times Z C Y is 1 times Z	You have $35.00. Each week (w) you spend $0.85. Which shows this? A $35.00 – $0.85(w) B. $0.85 / $35.00(w) C $35.00 = -$0.85(w)	Graciella is one year less than twice as old as her youngest brother. Which expression could be used to show her age? A 2b - 1 B 1 - 2b C. 2b + 1
special ed, kid 1	X		
special ed, kid 2		X	

Viewing #1 of 15

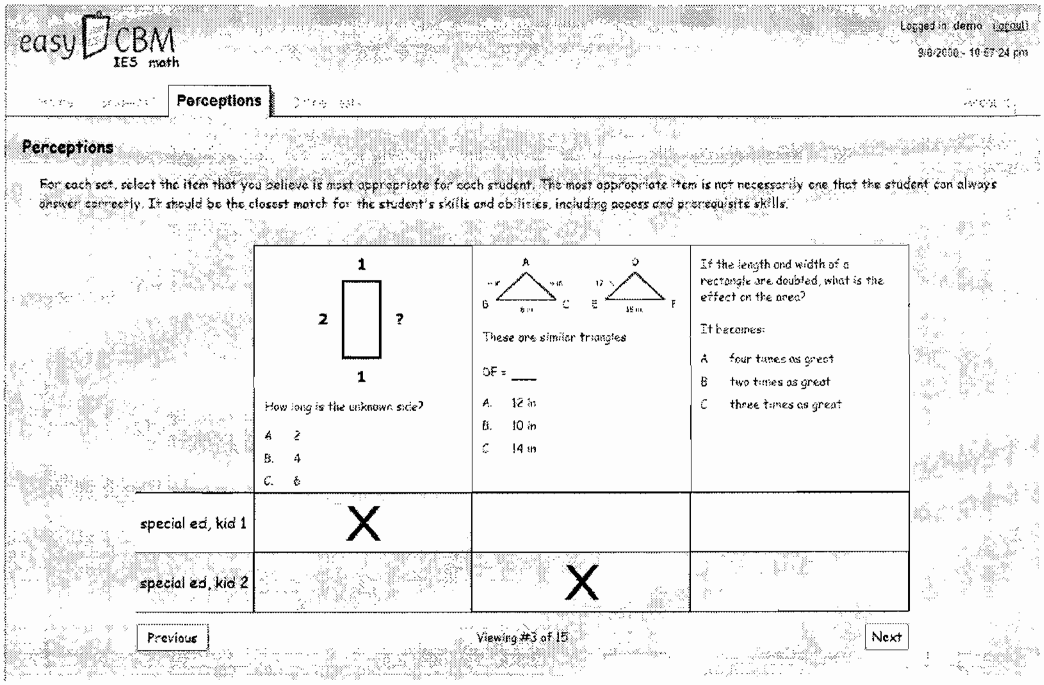


Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

	You have 10 candy bars. You give 50% away. How many candy bars did you give away? A 5 B 10 C 50	About 60% of the used white paper is recycled at Laura's school. The school uses 1,260 pounds of paper per month. Which is the best estimate for the number of pounds of white paper recycled per month? A 720-780 B 600-660 C 500-560
special ed, kid 1		X
special ed, kid 2	X	

Viewing #2 of 15

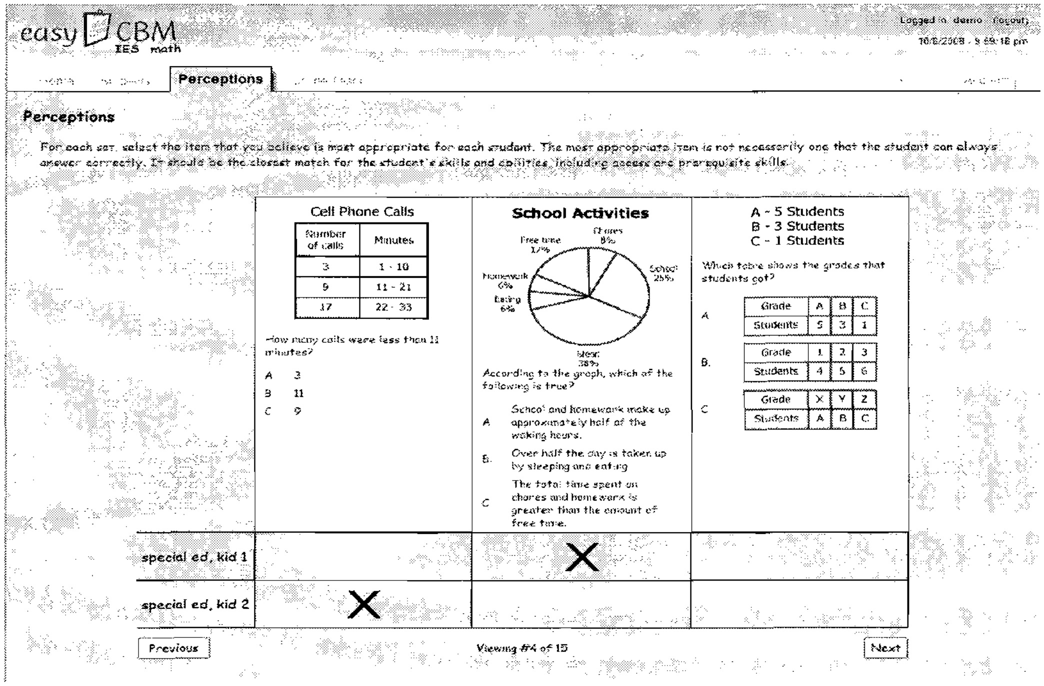


Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

	How long is the unknown side? A 2 B. 4 C. 6	These are similar triangles DF = ____ A. 12 in B. 10 in C 14 in	If the length and width of a rectangle are doubled, what is the effect on the area? It becomes: A four times as great B two times as great C three times as great
special ed, kid 1	X		
special ed, kid 2		X	

Viewing #3 of 15



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

Cell Phone Calls

Number of calls	Minutes
3	1 - 10
9	11 - 21
17	22 - 33

How many calls were less than 11 minutes?

A 3
B 11
C 9

School Activities

According to the graph, which of the following is true?

A School and homework make up approximately half of the waking hours.
B. Over half the day is taken up by sleeping and eating
C The total time spent on chores and homework is greater than the amount of free time.

A - 5 Students
B - 3 Students
C - 1 Students

Which table shows the grades that students got?

A

Grade	A	B	C
Students	5	3	1

B.

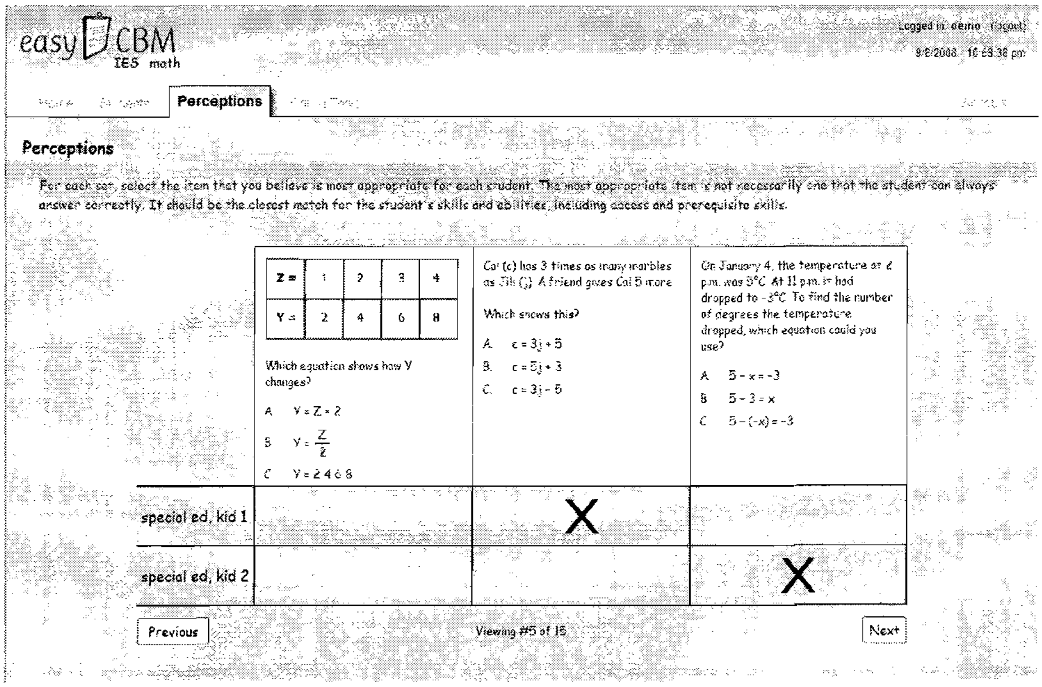
Grade	1	2	3
Students	4	5	6

C

Grade	X	Y	Z
Students	A	B	C

special ed, kid 1		X	
special ed, kid 2	X		

Viewing #4 of 15



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

Z =	1	2	3	4
Y =	2	4	6	8

Which equation shows how Y changes?

A $Y = Z \times 2$
B $Y = \frac{Z}{2}$
C Y = 2 4 6 8

Cal (c) has 3 times as many marbles as Jill (j). A friend gives Cal 5 more

Which shows this?

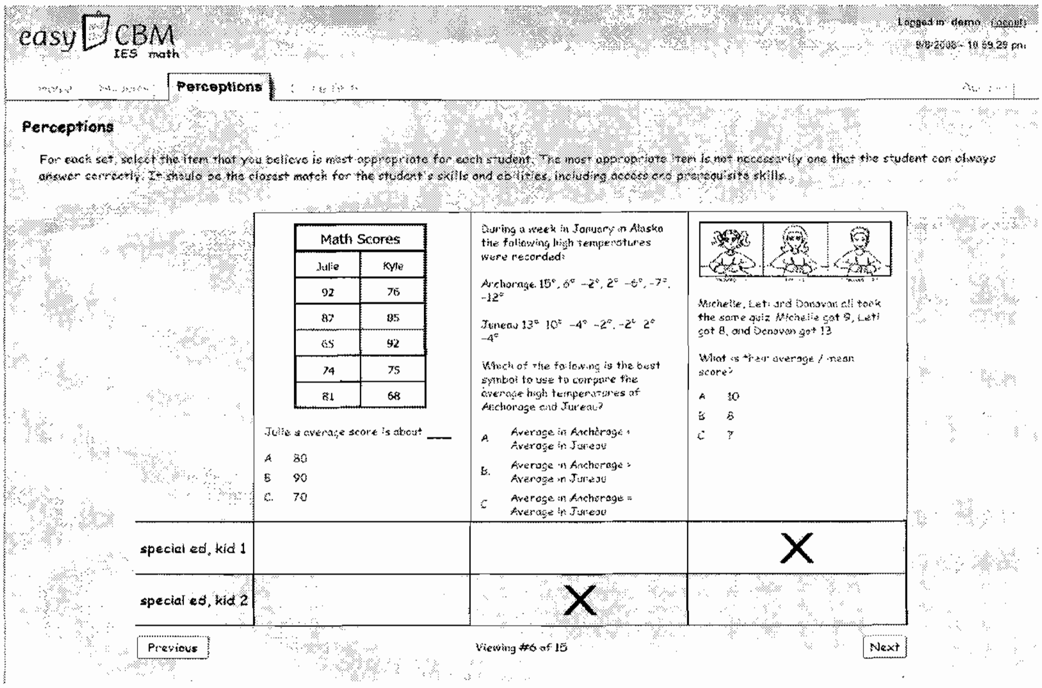
A. $c = 3j + 5$
B. $c = 5j + 3$
C. $c = 3j - 5$

On January 4, the temperature at 6 p.m. was 5°C. At 11 p.m. it had dropped to -3°C. To find the number of degrees the temperature dropped, which equation could you use?

A $5 - x = -3$
B $5 - 3 = x$
C $5 - (-x) = -3$

special ed, kid 1		X	
special ed, kid 2			X

Viewing #5 of 15



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

Math Scores	
Julie	Kyle
92	76
87	85
65	92
74	75
81	68

Julie s average score is about ____

A 80
B 90
C. 70

During a week in January in Alaska the following high temperatures were recorded:

Anchorage 15°, 6° −2°, 2° −6°, −7°, −12°

Juneau 13° 10° −4° −2°, −2° 2° −4°

Which of the following is the best symbol to use to compare the average high temperatures of Anchorage and Juneau?

A Average in Anchorage < Average in Juneau
B. Average in Anchorage > Average in Juneau
C Average in Anchorage = Average in Juneau

Michelle, Leti and Donovan all took the same quiz Michelle got 9, Leti got 8, and Donovan got 13

What is their average / mean score?

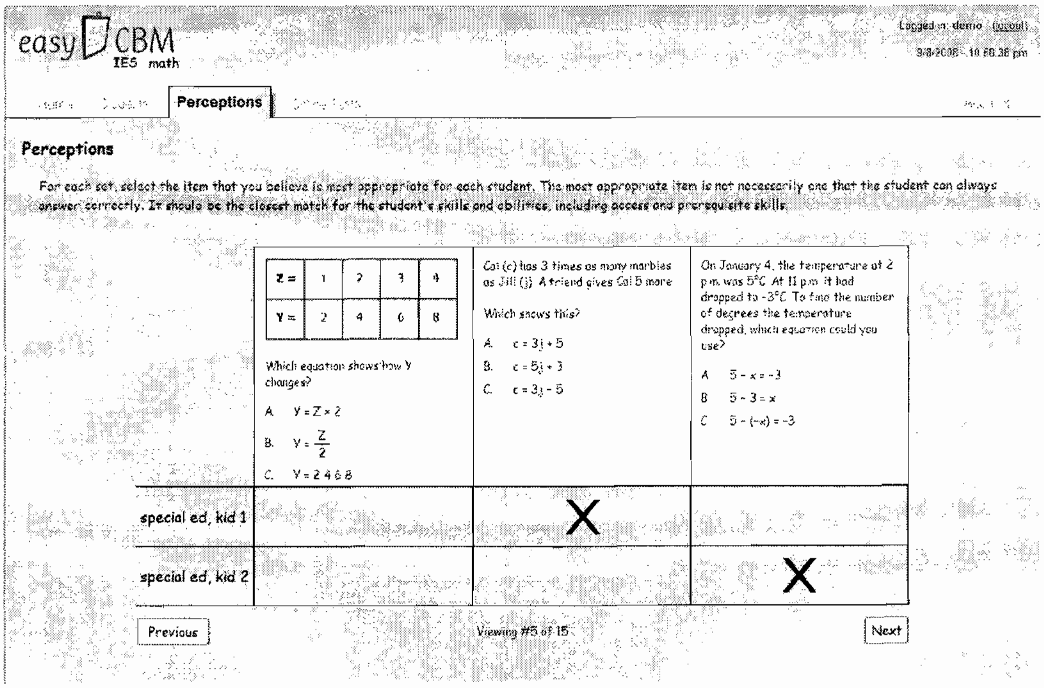
A 10
B 8
C 7

special ed, kid 1			X
special ed, kid 2		X	

Previous

Viewing #6 of 15

Next



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

Z =	1	2	3	4
Y =	2	4	6	8

Which equation shows how Y changes?

A. $Y = Z \times 2$
B. $Y = \frac{Z}{2}$
C. Y = 2 4 6 8

Cal (c) has 3 times as many marbles as Jill (j). A friend gives Cal 5 more

Which shows this?

A. $c = 3j + 5$
B. $c = 5j + 3$
C. $c = 3j - 5$

On January 4, the temperature at 2 p.m. was 5°C At 11 p.m. it had dropped to -3°C To find the number of degrees the temperature dropped, which equation could you use?

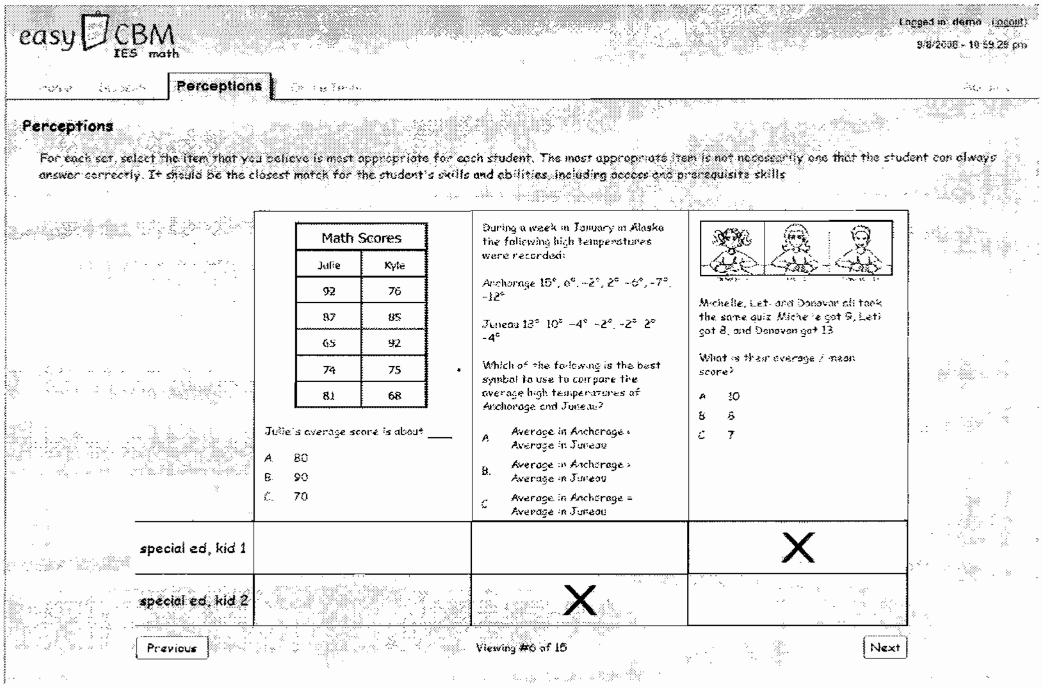
A $5 - x = -3$
B $5 - 3 = x$
C $5 - (-x) = -3$

special ed, kid 1		X	
special ed, kid 2			X

Previous

Viewing #5 of 15

Next



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills

Math Scores	
Julie	Kyle
92	76
87	85
65	92
74	75
81	68

Julie's average score is about ____

A. 80
B. 90
C. 70

During a week in January in Alaska the following high temperatures were recorded:

Anchorage 15°, 6°, -2°, 2° -6°, -7°, -12°

Juneau 13° 10° -4° -2°, -2° 2° -4°

Which of the following is the best symbol to use to compare the average high temperatures of Anchorage and Juneau?

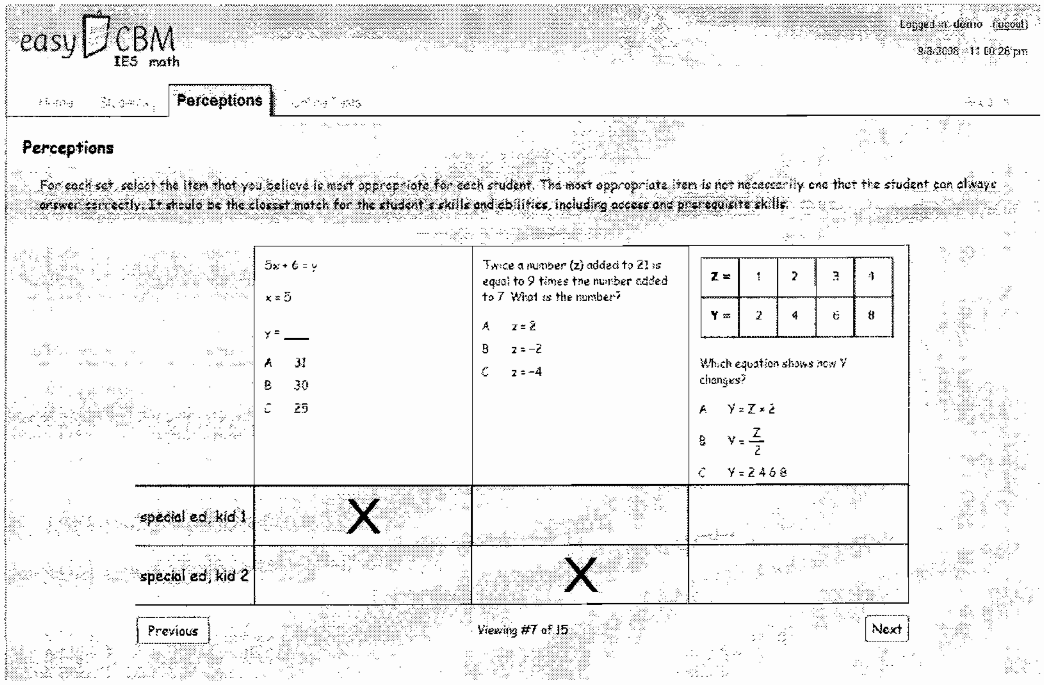
A. Average in Anchorage < Average in Juneau
B. Average in Anchorage > Average in Juneau
C. Average in Anchorage = Average in Juneau

Michelle, Leti and Donovan all took the same quiz. Michelle got 9, Leti got 8, and Donovan got 13

What is their average / mean score?

A 10
B 8
C 7

special ed, kid 1			X
special ed, kid 2		X	



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

$5x + 6 = y$

$x = 5$

$y =$ ____

A 31
B 30
C 25

Twice a number (z) added to 21 is equal to 9 times the number added to 7. What is the number?

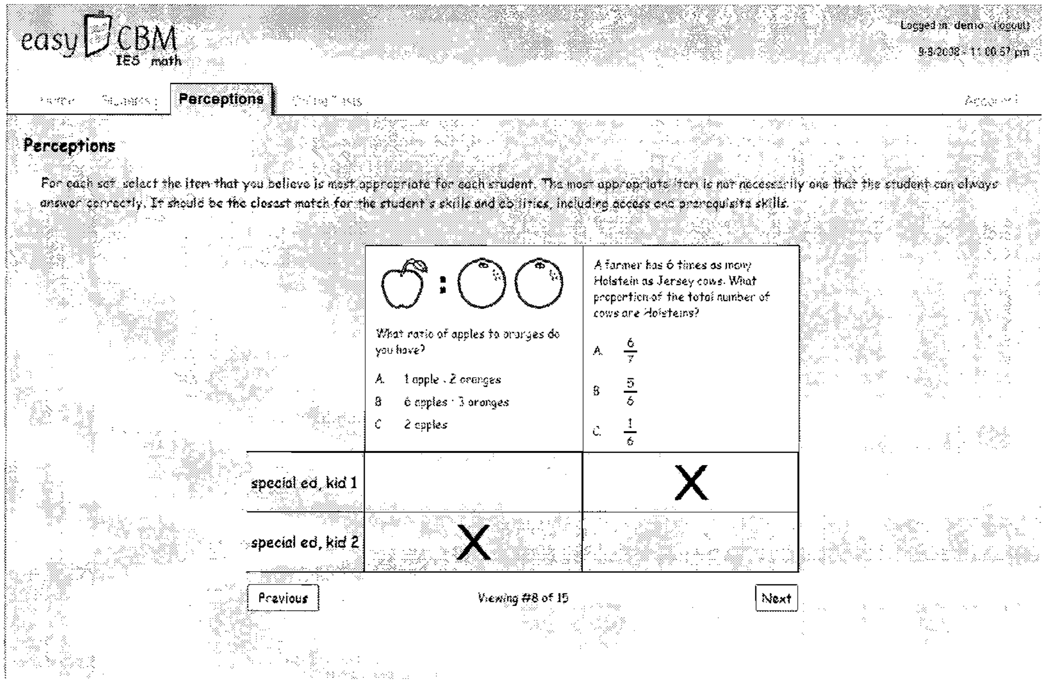
A $z = 2$
B $z = -2$
C $z = -4$

Z =	1	2	3	4
Y =	2	4	6	8

Which equation shows how Y changes?

A $Y = Z \times 2$
B $Y = \frac{Z}{2}$
C Y = 2 4 6 8

special ed, kid 1	X		
special ed, kid 2		X	

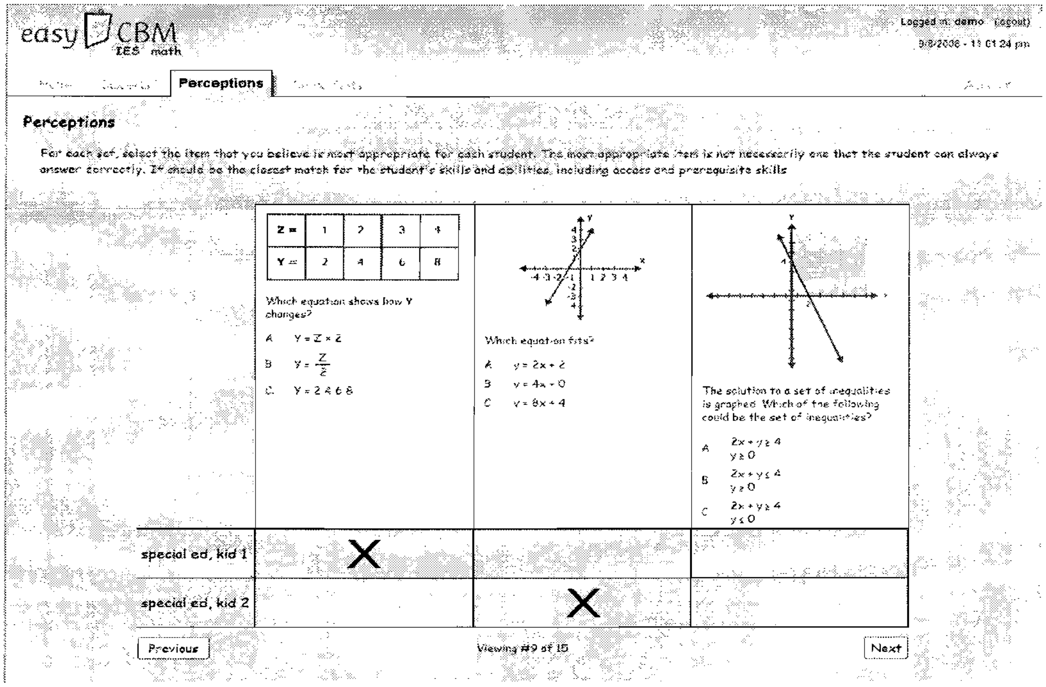


Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

	What ratio of apples to oranges do you have? A. 1 apple : 2 oranges B 6 apples : 3 oranges C 2 apples	A farmer has 6 times as many Holstein as Jersey cows. What proportion of the total number of cows are Holsteins? A. $\frac{6}{7}$ B $\frac{5}{6}$ C. $\frac{1}{6}$
special ed, kid 1		X
special ed, kid 2	X	

Viewing #8 of 15

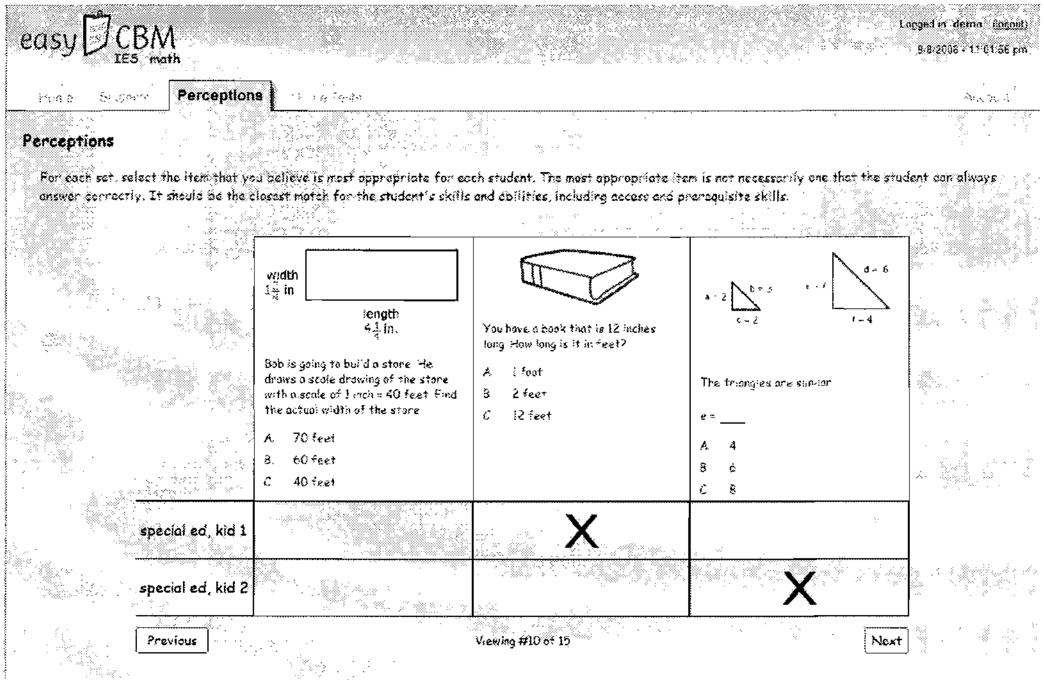


Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills

	Z = 1 2 3 4 Y = 2 4 6 8 Which equation shows how Y changes? A Y = Z × Z B $Y = \frac{Z}{2}$ C. Y = 2 4 6 8	Which equation fits? A $y = 2x + 2$ B $y = 4x + 0$ C $y = 8x + 4$	The solution to a set of inequalities is graphed. Which of the following could be the set of inequalities? A $2x + y \geq 4$ $y \geq 0$ B $2x + y \leq 4$ $y \geq 0$ C $2x + y \geq 4$ $y \leq 0$
special ed, kid 1	X		
special ed, kid 2		X	

Viewing #9 of 15

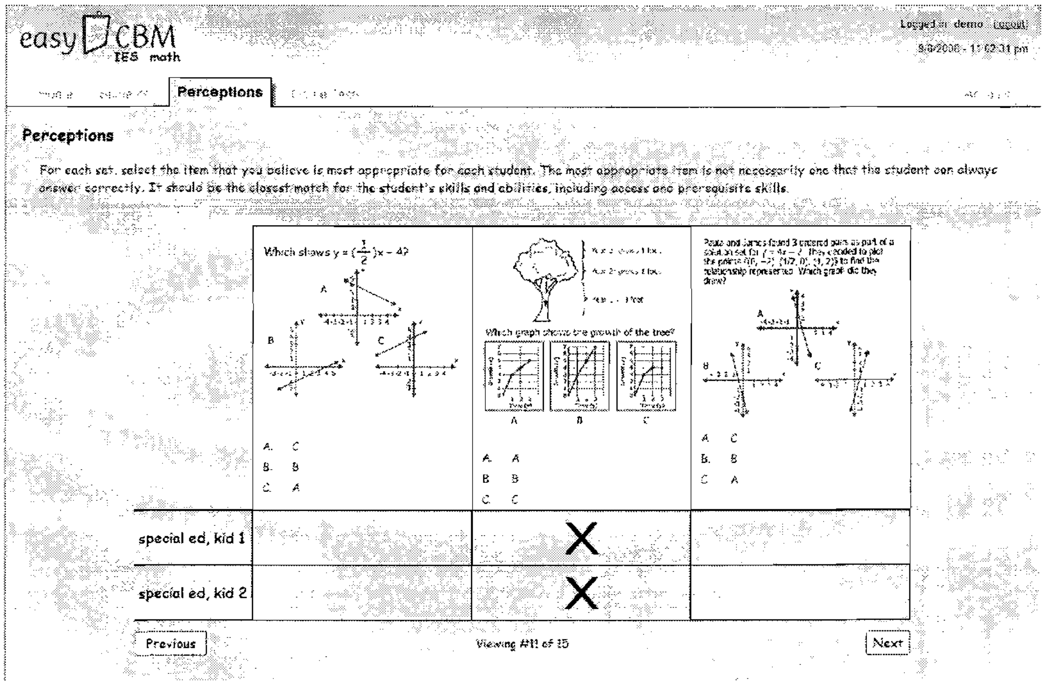


Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

	Bob is going to build a store. He draws a scale drawing of the store with a scale of 1 inch = 40 feet. Find the actual width of the store A. 70 feet B. 60 feet C. 40 feet	You have a book that is 12 inches long. How long is it in feet? A. 1 foot B. 2 feet C. 12 feet	The triangles are similar. e = ____ A. 4 B. 6 C. 8
special ed, kid 1		X	
special ed, kid 2			X

Viewing #10 of 15

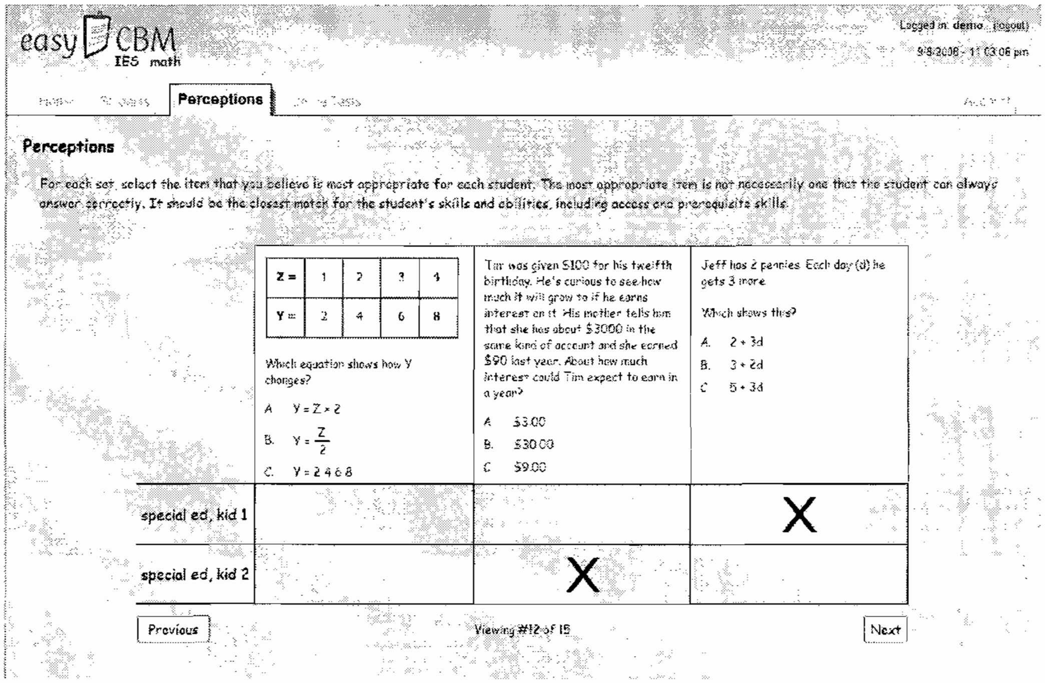


Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

	Which shows $y = (\frac{1}{2})x - 4$? A. C B. B C. A	Which graph shows the growth of the tree? A. A B. B C. C	Paulo and James found 3 ordered pairs as part of a solution set for $y = 4x - 2$. They decided to plot the points {(0, −2), (1/2, 0), (1, 2)} to find the relationship represented. Which graph did they draw? A. C B. B C. A
special ed, kid 1		X	
special ed, kid 2		X	

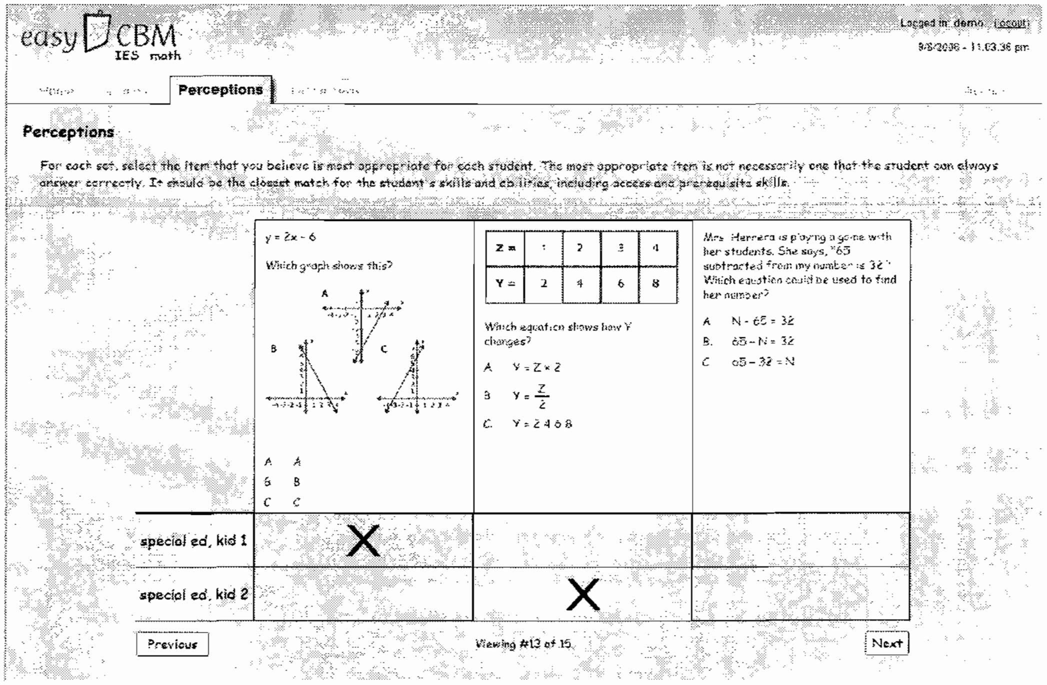
Viewing #11 of 15



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

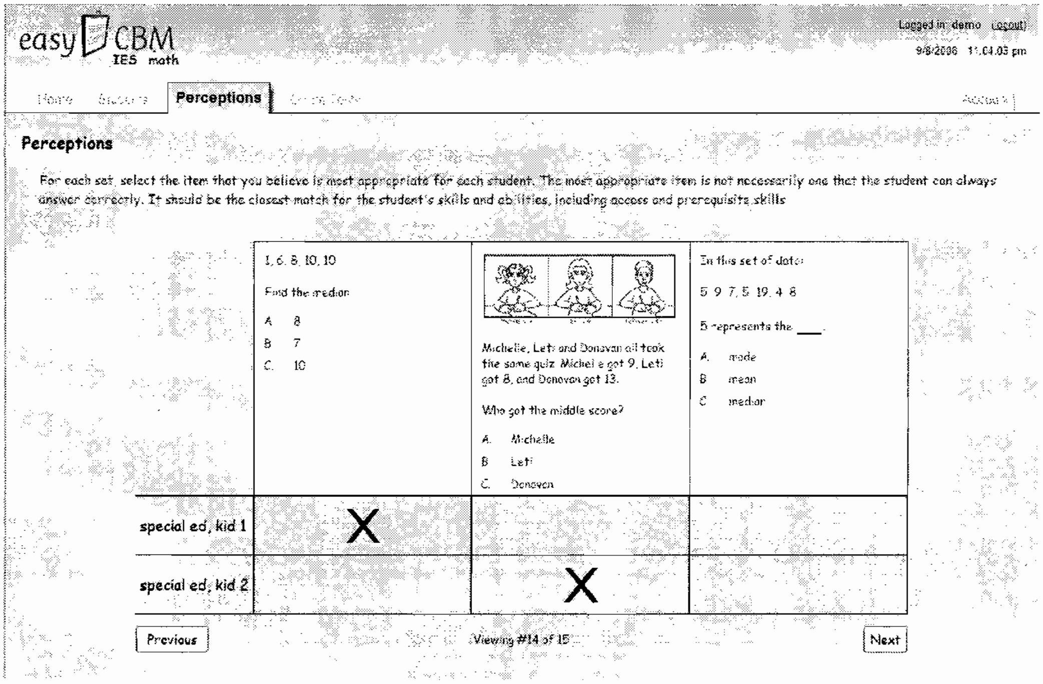
<table>
<tr><td></td><td><table><tr><td><b>Z =</b></td><td>1</td><td>2</td><td>3</td><td>4</td></tr><tr><td><b>Y =</b></td><td>2</td><td>4</td><td>6</td><td>8</td></tr></table>Which equation shows how Y changes?
A Y = Z × 2
B. $Y = \frac{Z}{2}$
C. Y = 2 4 6 8</td><td>Tim was given $100 for his twelfth birthday. He's curious to see how much it will grow to if he earns interest on it. His mother tells him that she has about $3000 in the same kind of account and she earned $90 last year. About how much interest could Tim expect to earn in a year?
A $3.00
B. $30.00
C $9.00</td><td>Jeff has 2 pennies. Each day (d) he gets 3 more.
Which shows this?
A. 2 + 3d
B. 3 + 2d
C 5 + 3d</td></tr>
<tr><td><b>special ed, kid 1</b></td><td></td><td></td><td>X</td></tr>
<tr><td><b>special ed, kid 2</b></td><td></td><td>X</td><td></td></tr>
</table>



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

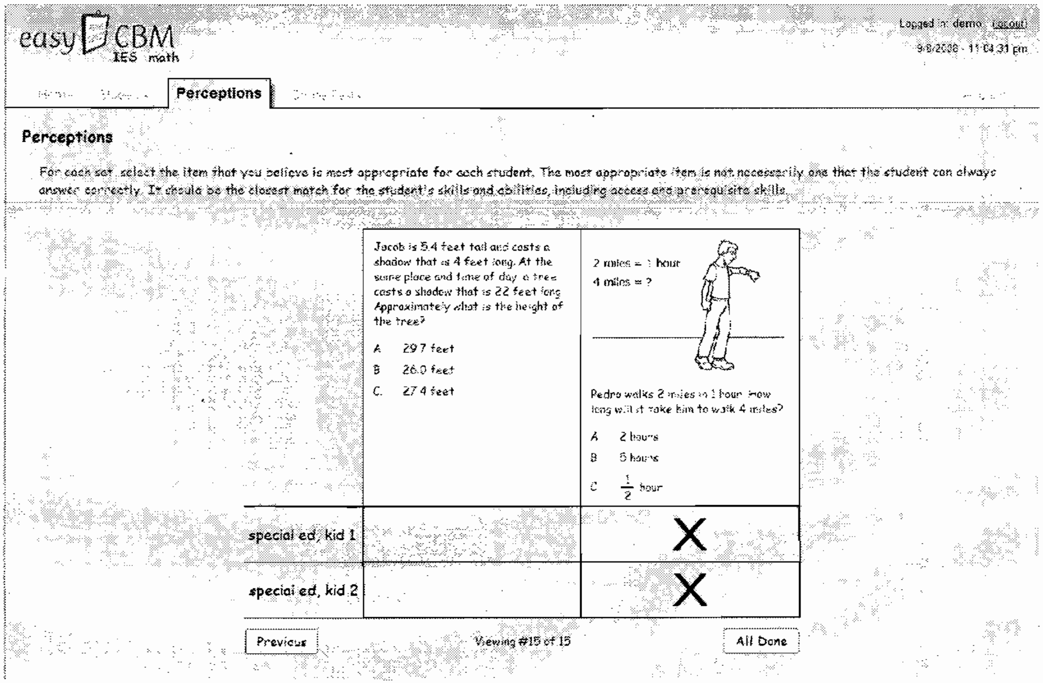
<table>
<tr><td></td><td>y = 2x - 6
Which graph shows this?
A A
B B
C C</td><td><table><tr><td><b>Z =</b></td><td>1</td><td>2</td><td>3</td><td>4</td></tr><tr><td><b>Y =</b></td><td>2</td><td>4</td><td>6</td><td>8</td></tr></table>Which equation shows how Y changes?
A Y = Z × 2
B $Y = \frac{Z}{2}$
C. Y = 2 4 6 8</td><td>Mrs. Herrera is playing a game with her students. She says, "65 subtracted from my number is 32." Which equation could be used to find her number?
A N - 65 = 32
B. 65 – N = 32
C 65 – 32 = N</td></tr>
<tr><td><b>special ed, kid 1</b></td><td>X</td><td></td><td></td></tr>
<tr><td><b>special ed, kid 2</b></td><td></td><td>X</td><td></td></tr>
</table>



Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills

	1, 6, 8, 10, 10 Find the median A 8 B 7 C. 10	Michelle, Leti and Donavan all took the same quiz. Michelle got 9, Leti got 8, and Donovan got 13. Who got the middle score? A. Michelle B Leti C. Donovan	In this set of data: 5 9 7, 5 19, 4 8 5 represents the ___. A. mode B mean C median
special ed, kid 1	X		
special ed, kid 2		X	



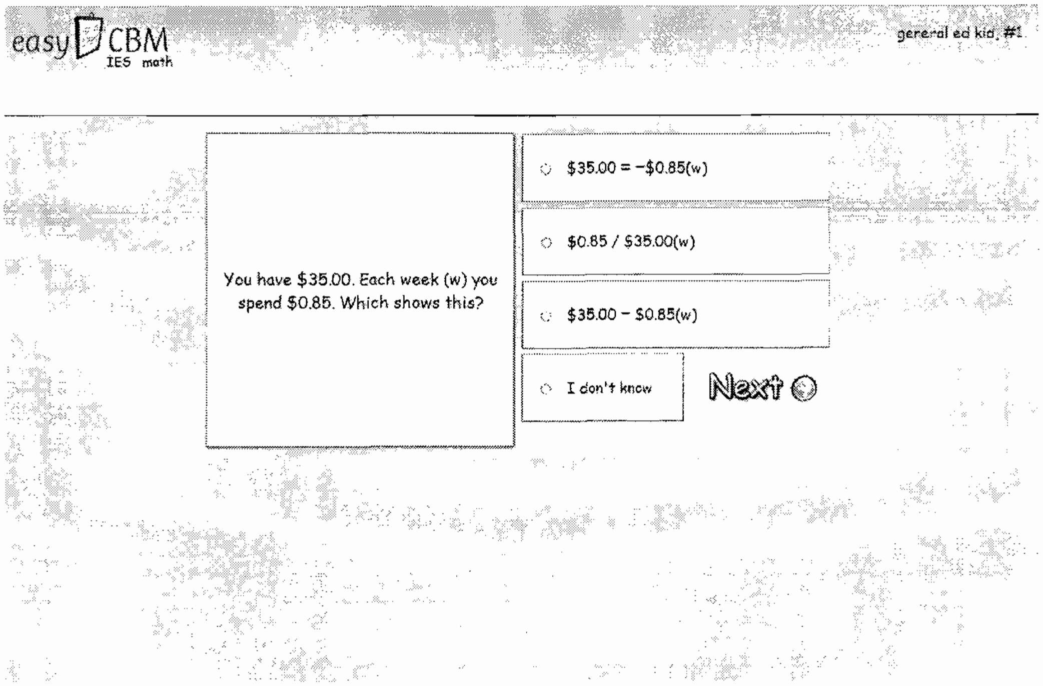
Perceptions

For each set, select the item that you believe is most appropriate for each student. The most appropriate item is not necessarily one that the student can always answer correctly. It should be the closest match for the student's skills and abilities, including access and prerequisite skills.

	Jacob is 5.4 feet tall and casts a shadow that is 4 feet long. At the same place and time of day, a tree casts a shadow that is 22 feet long. Approximately what is the height of the tree? A 29.7 feet B 26.0 feet C. 27.4 feet	2 miles = 1 hour 4 miles = ? Pedro walks 2 miles in 1 hour. How long will it take him to walk 4 miles? A 2 hours B 5 hours C $\frac{1}{2}$ hour
special ed, kid 1		X
special ed, kid 2		X

APPENDIX D

SAMPLE OF STUDENT PERFORMANCE TEST



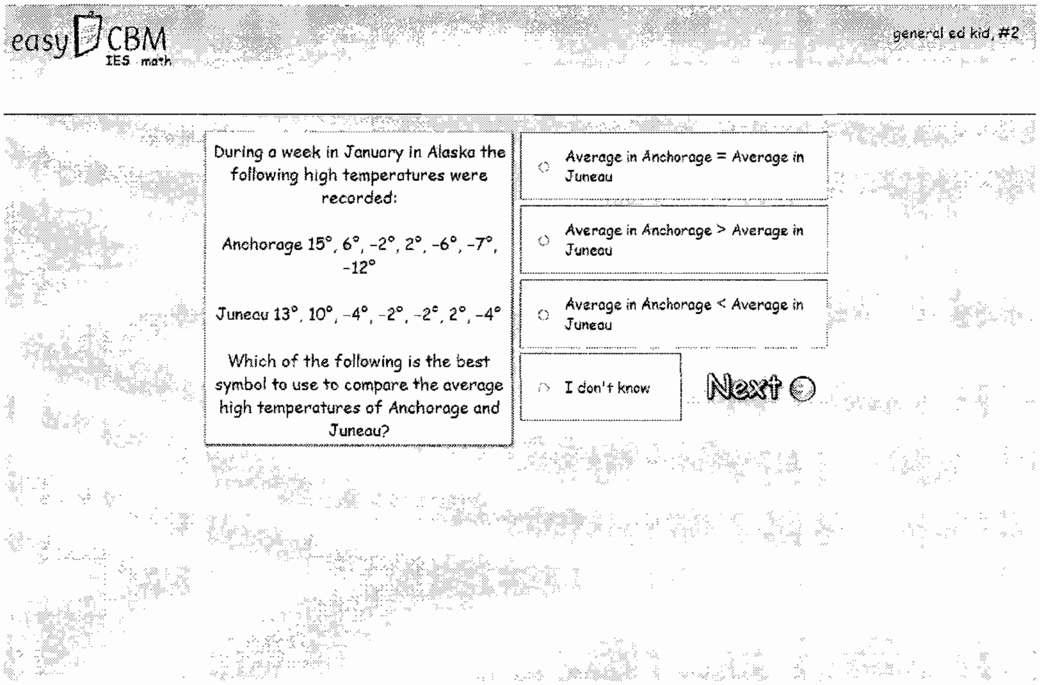
You have $35.00. Each week (w) you spend $0.85. Which shows this?

○ $35.00 = −$0.85(w)

○ $0.85 / $35.00(w)

○ $35.00 − $0.85(w)

○ I don't know



general ed kid, #2

During a week in January in Alaska the following high temperatures were recorded:

Anchorage 15°, 6°, -2°, 2°, -6°, -7°, -12°

Juneau 13°, 10°, -4°, -2°, -2°, 2°, -4°

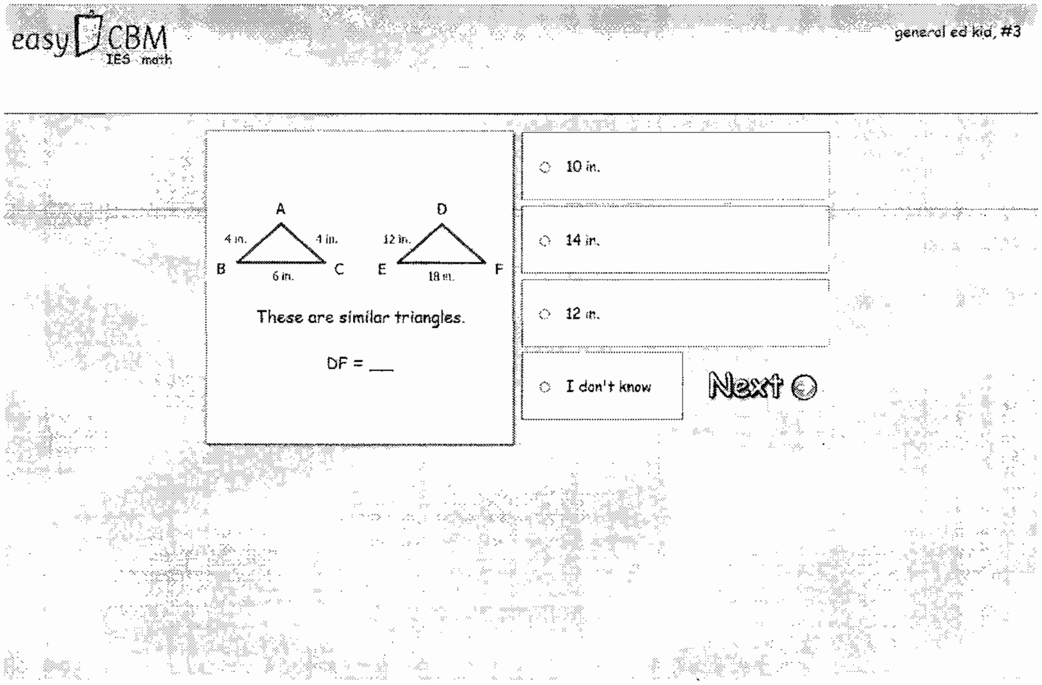
Which of the following is the best symbol to use to compare the average high temperatures of Anchorage and Juneau?

○ Average in Anchorage = Average in Juneau

○ Average in Anchorage > Average in Juneau

○ Average in Anchorage < Average in Juneau

○ I don't know



general ed kid, #3

These are similar triangles.

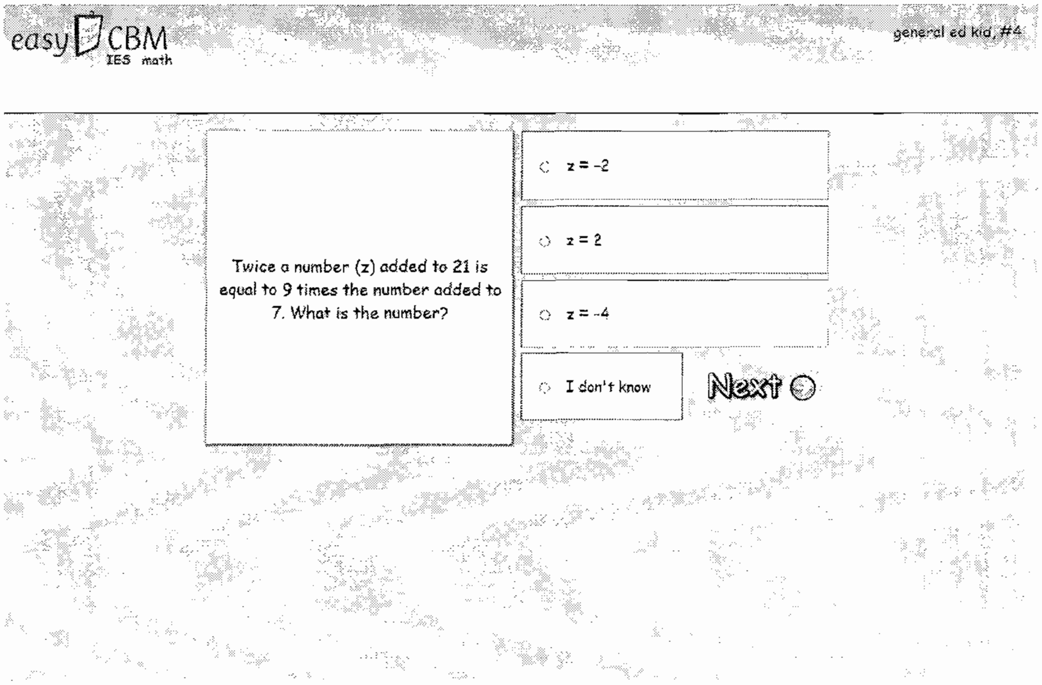
DF = ___

○ 10 in.

○ 14 in.

○ 12 in.

○ I don't know



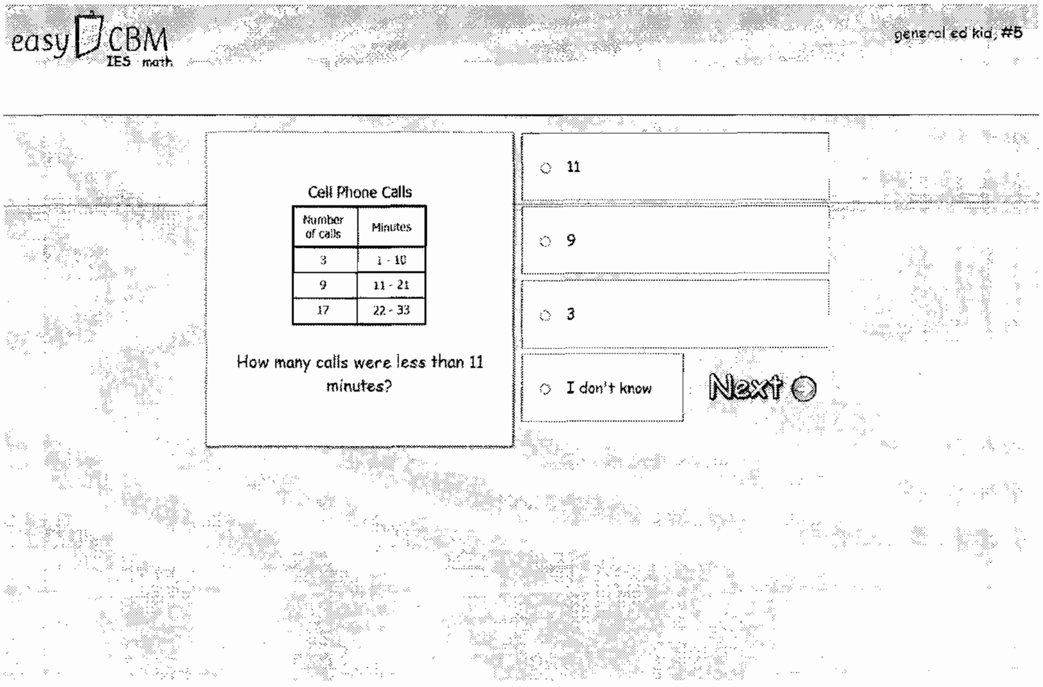
Twice a number (z) added to 21 is equal to 9 times the number added to 7. What is the number?

○ z = -2

○ z = 2

○ z = -4

○ I don't know



Cell Phone Calls

Number of calls	Minutes
3	1 - 10
9	11 - 21
17	22 - 33

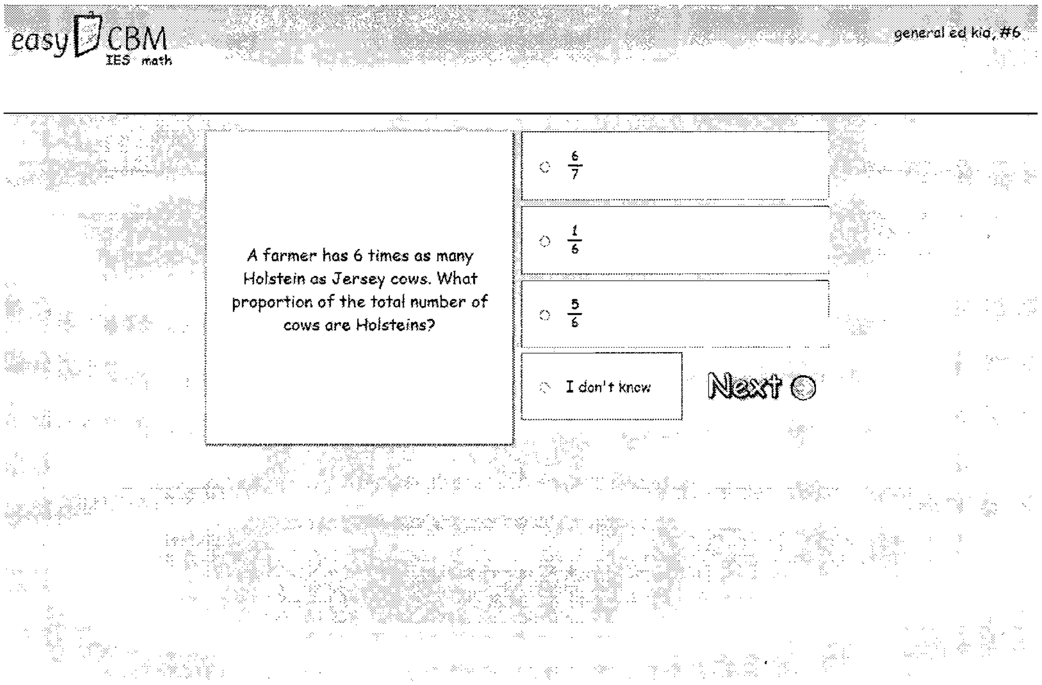
How many calls were less than 11 minutes?

○ 11

○ 9

○ 3

○ I don't know



easy CBM IES math

general ed kid, #6

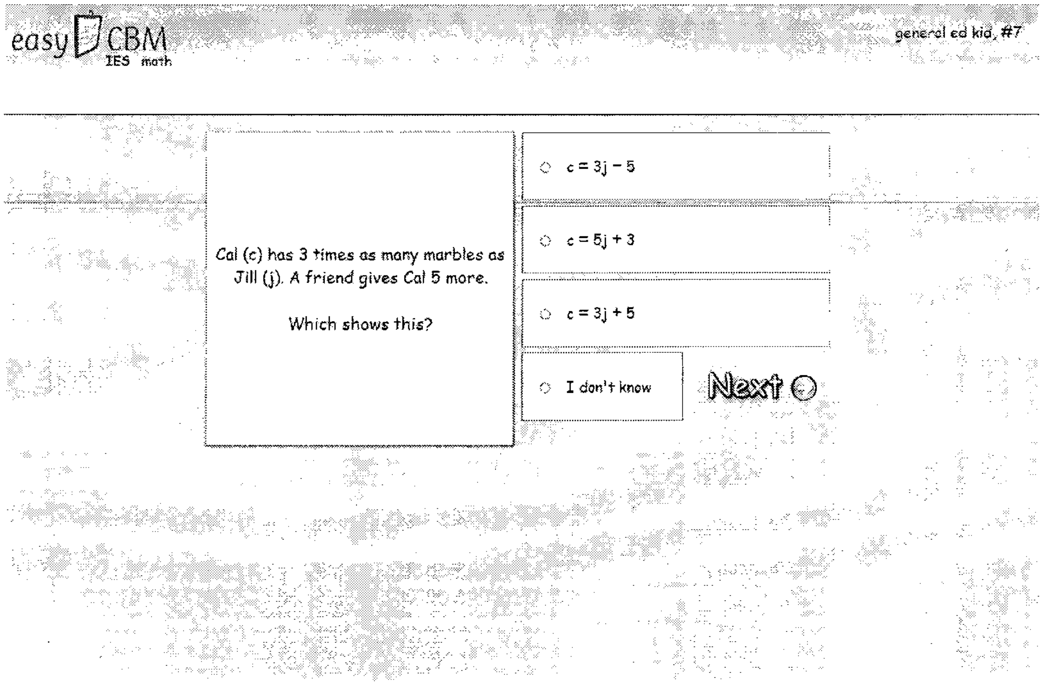
A farmer has 6 times as many Holstein as Jersey cows. What proportion of the total number of cows are Holsteins?

○ $\frac{6}{7}$

○ $\frac{1}{6}$

○ $\frac{5}{6}$

○ I don't know



easy CBM IES math

general ed kid, #7

Cal (c) has 3 times as many marbles as Jill (j). A friend gives Cal 5 more.

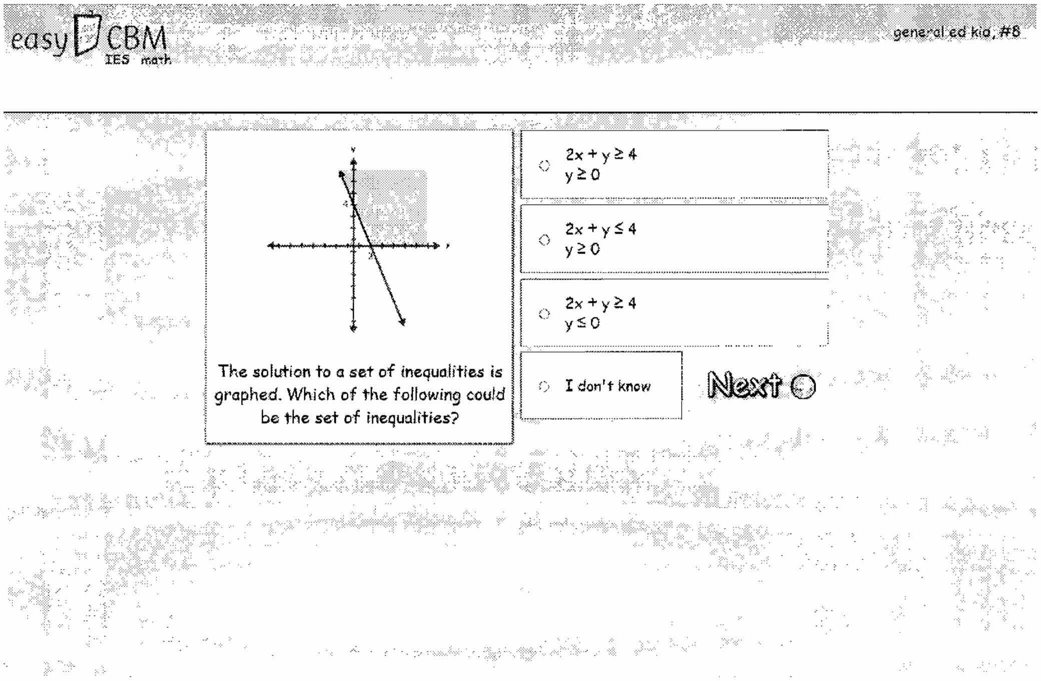
Which shows this?

○ $c = 3j - 5$

○ $c = 5j + 3$

○ $c = 3j + 5$

○ I don't know



general ed kid, #8

The solution to a set of inequalities is graphed. Which of the following could be the set of inequalities?

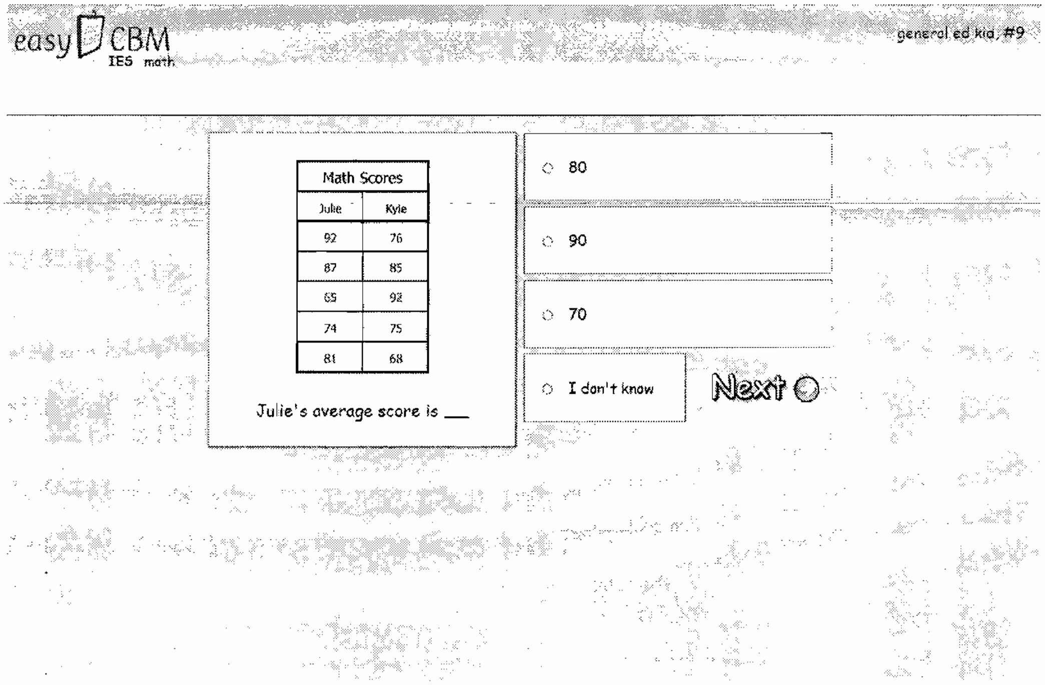
○ $2x + y \geq 4$
$y \geq 0$

○ $2x + y \leq 4$
$y \geq 0$

○ $2x + y \geq 4$
$y \leq 0$

○ I don't know

Next ○



general ed kid, #9

Math Scores	
Julie	Kyle
92	76
87	85
65	92
74	75
81	68

Julie's average score is ___

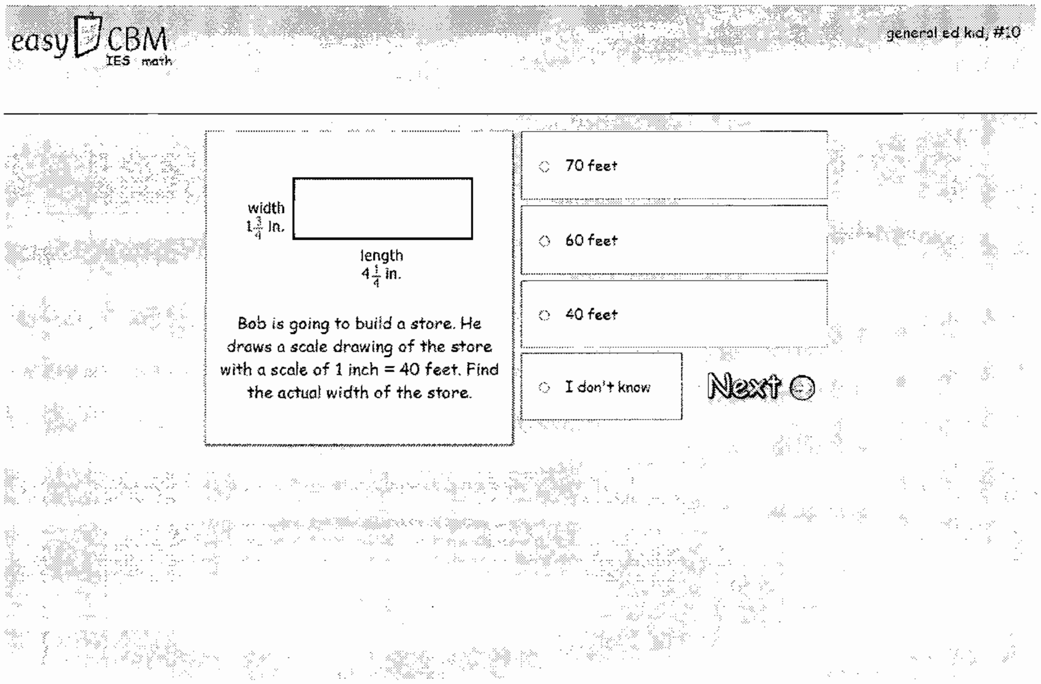
○ 80

○ 90

○ 70

○ I don't know

Next ○



general ed kid, #10

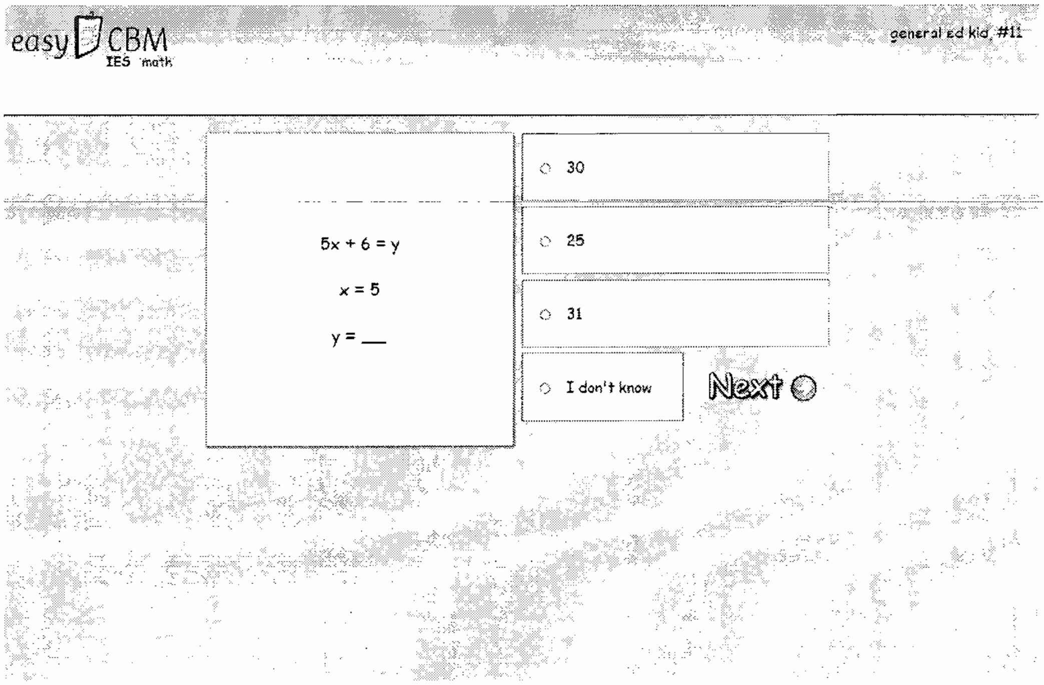
Bob is going to build a store. He draws a scale drawing of the store with a scale of 1 inch = 40 feet. Find the actual width of the store.

○ 70 feet

○ 60 feet

○ 40 feet

○ I don't know



general ed kid, #11

$5x + 6 = y$

$x = 5$

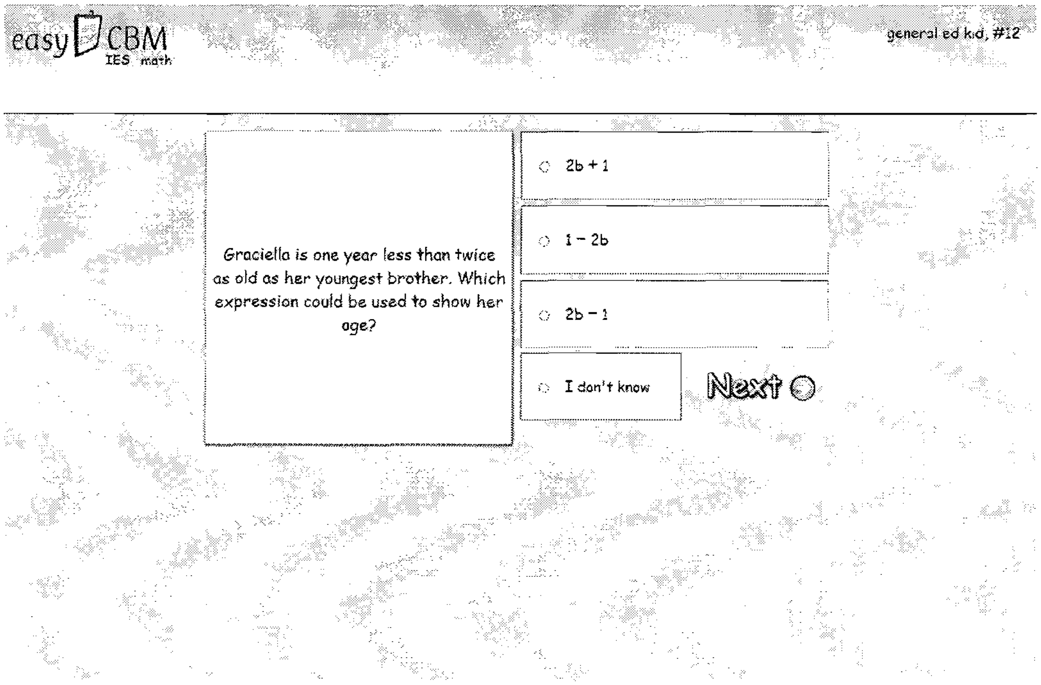
$y =$ __

○ 30

○ 25

○ 31

○ I don't know



general ed kid, #12

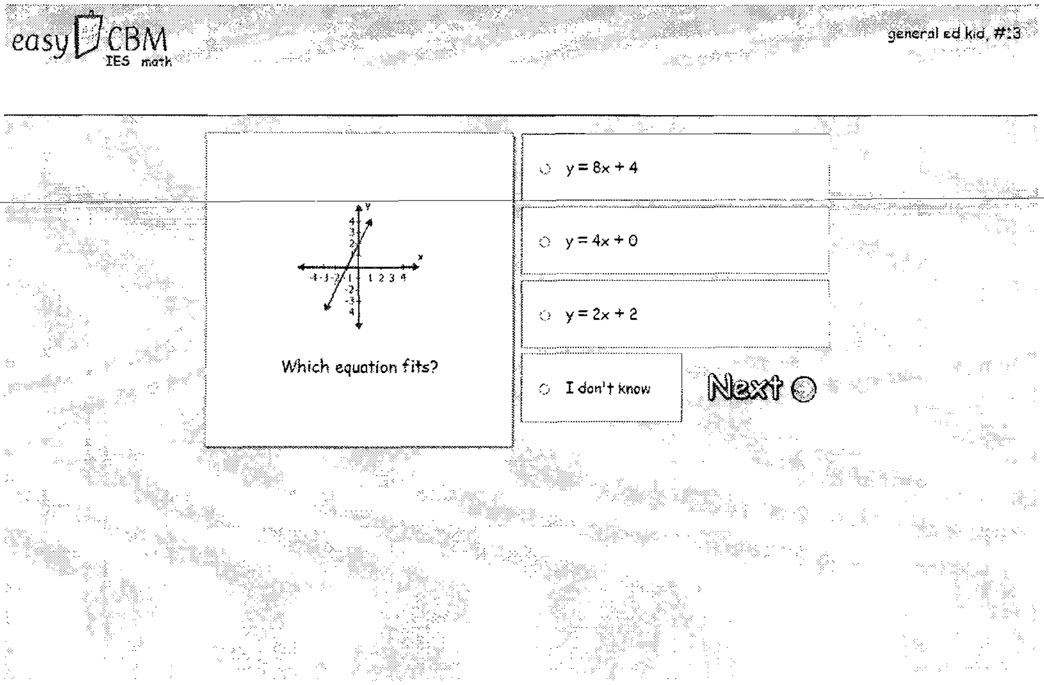
Graciella is one year less than twice as old as her youngest brother. Which expression could be used to show her age?

○ 2b + 1

○ 1 − 2b

○ 2b − 1

○ I don't know



general ed kid, #13

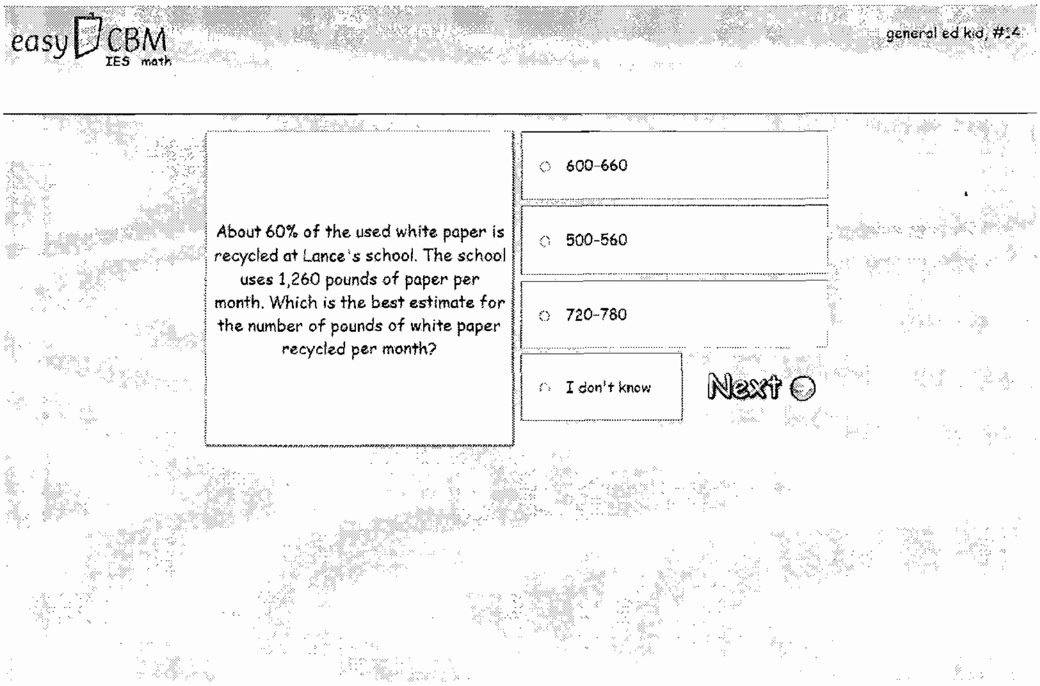
Which equation fits?

○ y = 8x + 4

○ y = 4x + 0

○ y = 2x + 2

○ I don't know



general ed kid, #14

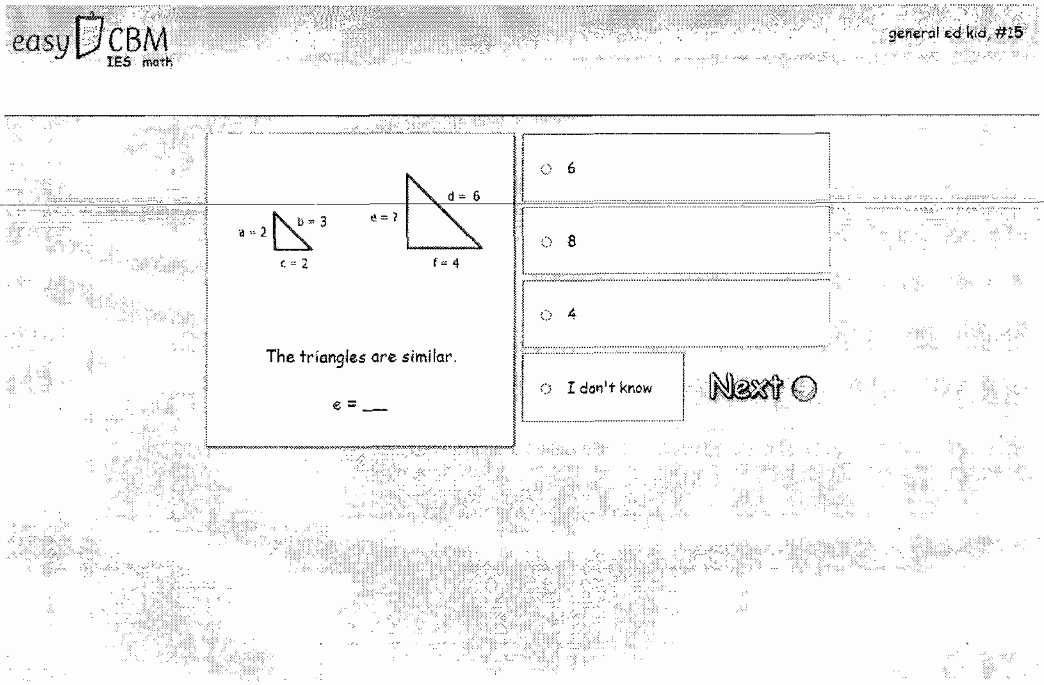
About 60% of the used white paper is recycled at Lance's school. The school uses 1,260 pounds of paper per month. Which is the best estimate for the number of pounds of white paper recycled per month?

○ 600-660

○ 500-560

○ 720-780

○ I don't know



general ed kid, #15

The triangles are similar.

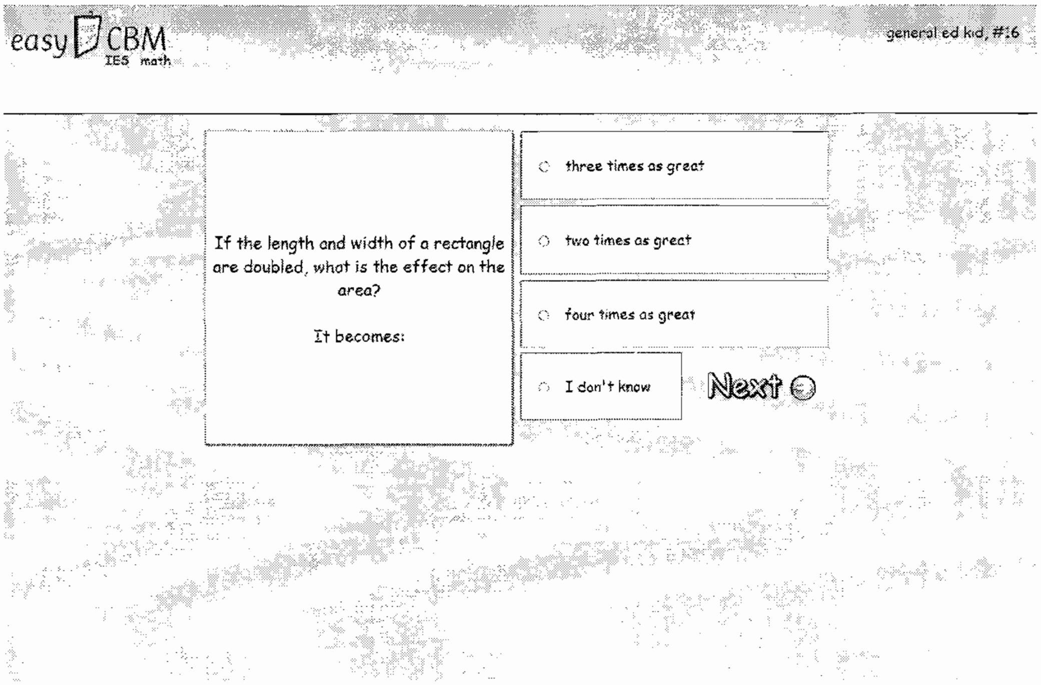
e = __

○ 6

○ 8

○ 4

○ I don't know



easy CBM IES math

general ed kid, #16

If the length and width of a rectangle are doubled, what is the effect on the area?

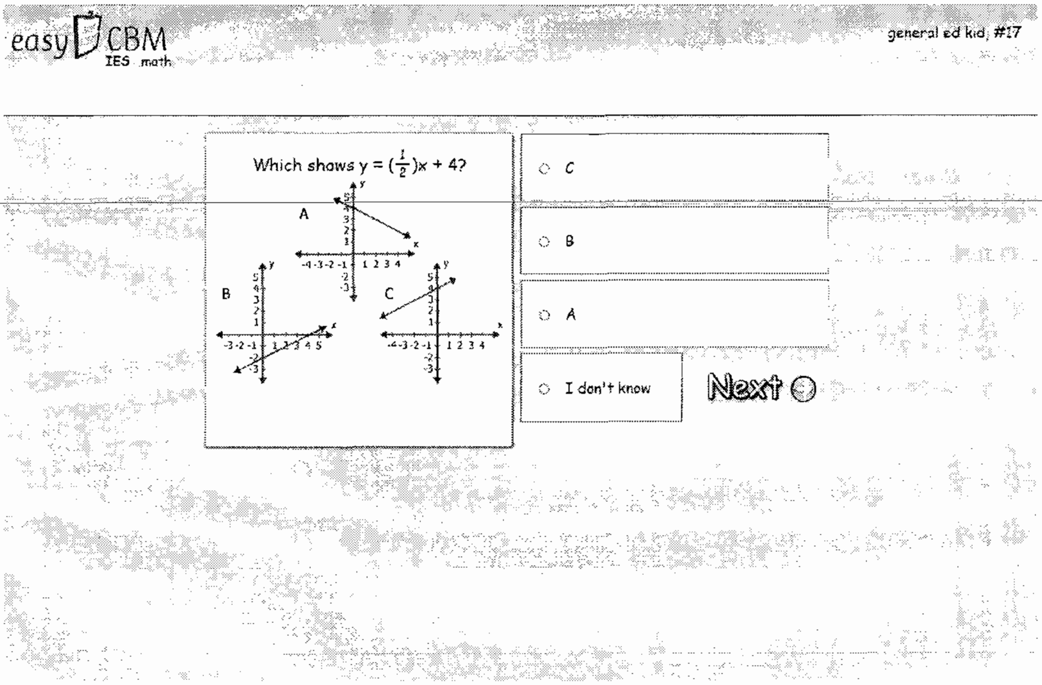
It becomes:

○ three times as great

○ two times as great

○ four times as great

○ I don't know



easy CBM IES math

general ed kid, #17

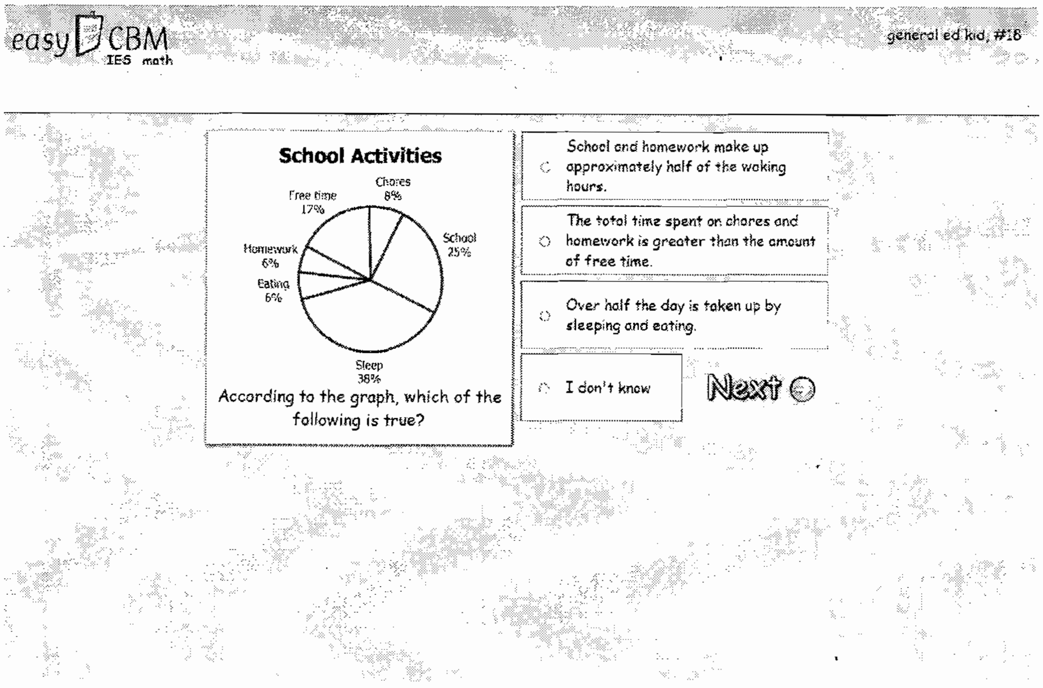
Which shows $y = (\frac{1}{2})x + 4$?

○ C

○ B

○ A

○ I don't know



general ed kid, #18

School Activities

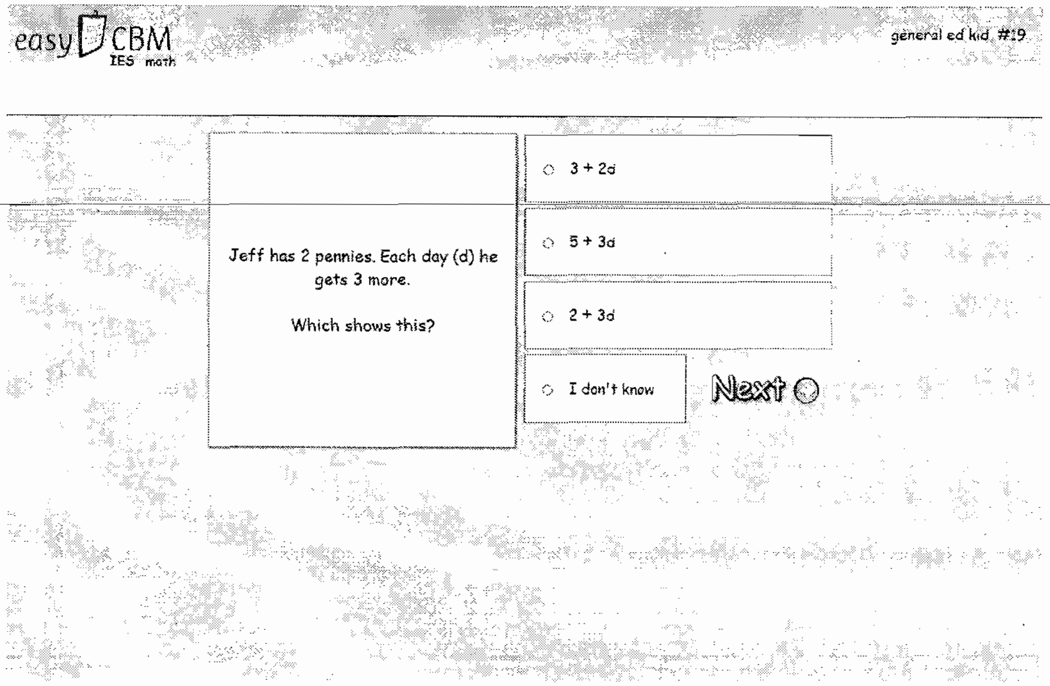
According to the graph, which of the following is true?

○ School and homework make up approximately half of the waking hours.

○ The total time spent on chores and homework is greater than the amount of free time.

○ Over half the day is taken up by sleeping and eating.

○ I don't know



general ed kid, #19

Jeff has 2 pennies. Each day (d) he gets 3 more.

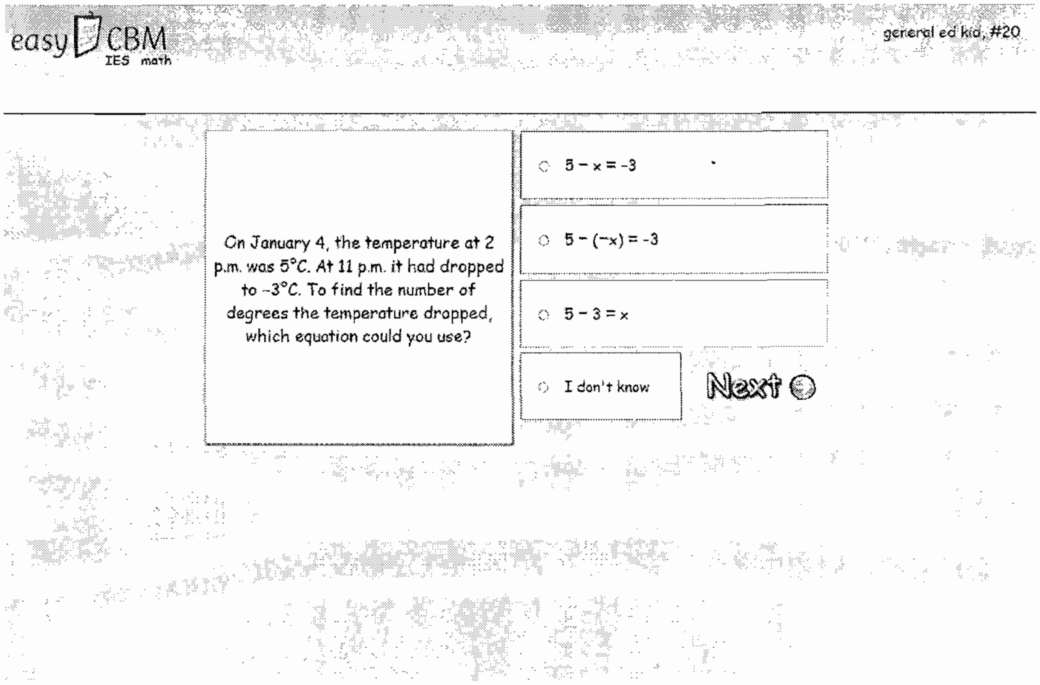
Which shows this?

○ 3 + 2d

○ 5 + 3d

○ 2 + 3d

○ I don't know



general ed kid, #20

On January 4, the temperature at 2 p.m. was 5°C. At 11 p.m. it had dropped to −3°C. To find the number of degrees the temperature dropped, which equation could you use?

○ $5 - x = -3$

○ $5 - (-x) = -3$

○ $5 - 3 = x$

○ I don't know

REFERENCES

Almond P. J., Lehr, C., Thurlow, M. L., & Quenemoen, R. (2002). Participation in large-scale state assessment and accountability systems. In J. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 341-370). Mahwah, NJ: Erlbaum.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In Paul. H, & Wainer, H (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Erlbaum.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and practice, 22*(3), 21-29.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals.* New York: David McKay.

Browder, D. M., & Spooner, F. (Eds.) (2006). Teaching language arts, math, and science to students with significant cognitive disabilities. Baltimore, MD: Paul H. Brookes.

Burling, K. (2007a). *NCLB regulations for alternate achievement standards (1%): A white paper from Pearson Educational Measurement.* Retrieved January 15, 2008 from http://www.pearsonedmeasurement.com/downloads/white/wp0701.pdf

Burling, K. (2007b). *NCLB regulations for modified achievement standards (2%): A white paper from Pearson Educational Measurement.* Retrieved January 15, 2008 from http://www.pearsonsolutions.com/downloads/white/wp0702.pdf

California Department of Education (2008). *California modified assessment (CMA).* Retrieved April 1, 2008 from http://www.cde.ca.gov/ta/tg/sr/cmastar.asp

Clauser, B. F., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Crane, P. K., Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine. 23*, 241-256.

Embreston, S. E., Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement Issues and Practices, 24*, 3-14.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21-35.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*(5), 394-411.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17-27.

Haladyna, T. M. (2002). Supporting documentation: Assuring more valid test score interpretations and uses. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 89-108). Mahwah, NJ: Erlbaum.

Holland, P. W., & Thayer, D. P. (1988). Differential item functioning and the Mantel-Haenszel prodecure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Individuals with Disabilities Education Act Amendments of 1997, PL 105-17, 20 U. S. C., §§ 1400 *et seq.*

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory into Practice, 42, 212-218.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*(3), 6-14.

Lazarus, S. S., Thurlow, M. L., Christensen, L. L., & Cormier, D.(2007). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2007* (Synthesis Report 67). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Linacre, J. M. (2006). WINSTEPS Rasch measurement computer program. Chicago: Winsteps. com.

Marion. S. (2007, July 26). *A technical design and documentation workbook for assessments based on modified achievement standards.* Minneapolis MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 28, 2008 from http://cehd.umn.edu/NCEO/Teleconferences/AAMASteleconferences/AAMASworkbook.pdf.

Mazor, K.M., Clauser, B.E., & Hambleton, R.K. (1992). The effect on sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.

McDaniel, E., & Lawrence, C. (1990). *Levels of cognitive complexity: An approach to the measurement of thinking.* New York: Springer-Verlag.

Mellenberg. G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-108.

Messick, S. (1994). The interplay of evidence and consequences in the validity of performance assessments. *Educational Researcher, 23*(2), 13-23.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115-135.

No Child Left Behind Act of 2001, Pub. L. No. 107-110. 115 Stat. 1425 (2002).

Oklahoma State Department of Education (2006). *Oklahoma modified assessment program (OMAAP)*. Oklahoma City: Oklahoma State of Education, Office of Accountability and Assessment and Special Education Services. Retrieved April 3, 2008 from http://www.sde.state.ok.us/home/defaultie.html.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32*, 303-316.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Rabinowitz, S. N., & Sato, E. (2006). *The technical adequacy of assessments for alternate student populations.* San Francisco: WestEd. Assessment and Accountability Comprehensive Center. Retrieved April 9, 2008 from http://www.aacompcenter.org/pdf/taasp.pdf

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment in special and inclusive education.* Boston: Houghton Mifflin.

Sato, E., Rabinowitz, S., Worth, P., Gallagher, C., Lagunoff, R., & Crane, E. (2007). *Evaluation of the technical evidence of assessments for special student populations.* (Assessment and Accountability Comprehensive Center report). San Francisco: WestEd.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistics regression procedures. *Journal of Educational Measurement, 27*, p. 361-370.

Thurlow, M.L., Elliott, J.L., & Ysseldyke, J. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements.* Thousand Oaks, CA: Corwin Press.

Thurlow, M.L., Bielinski, J., Minnema, J., & Scott, J. (2002). Out-of-level testing revisited: new concerns in the era of standards-based reform. In J. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* Mahwah, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118-128.

Tindal, G., McDonald, M., Tedesco, M., Glasgow, A., Almond, P., Crawford, L., & Hollenbeck, K. (2003). Alternate assessment in reading and math: development and validation for students with significant disabilities. *Exceptional Children, 69*(4), p. 481-494.

U.S. Department of Education. Title I- Improving the Academic Achievement of the Disadvantaged Final Rule. 68 Fed. Reg. (Dec. 9, 2003).

U.S. Department of Education. Title I- Improving the Academic Achievement of the Disadvantaged Final Rule. 72 Fed. Reg. (Apr. 9, 2007).

U.S. Office of Special Education Programs (2006). *Validity evidence.* Retrieved February 15, 2008 from http://www.osepideasthatwork.org/toolkit/tk_validityevidnce.asp.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states.* Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states.* Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25.

Yovanoff, P., & Tindal, G. (2007). Scaling early reading alternate assessments with statewide measures. *Exceptional Children, 73*(2), 184-201.

Zheng, Y., Gierl, M. J., & Cui, Y. (2007, April). *Using Real Data to Compare DIF Detection and Effect Size Measures among Mantel-Haenszel, SIBTEST, and Logistic Regression Procedures*. Paper presented at the Annual National Council on Measurement in Education (NCME) conference, Chicago, IL.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.